NGS Data Analysis Workshop

Course Tutorial

For more information please contact:

Genomatix Software GmbH
Bayerstr. 85a
80335 Munich
Germany

Phone: +49 89 599766 0
Fax: +49 89 599766 55
Email: info@genomatix.de
WWW: http://www.genomatix.de

genomatix

Table of Contents

Introduction

Next Generation Sequencing (NGS) offers a sensitive and unbiased method for high-throughput genomic studies. NGS is complementing, and to a considerable extent supplanting longer established methods, such as microarrays, in the analysis of e.g. gene expression, protein-DNA binding, or chromatin modification on a genome-wide scale.

A number of suppliers offer platforms for massive parallel sequencing. Throughput grows with each new sequencer generation, and with increasing numbers of reads per experiment, the scalability of the mapping algorithm is becoming an important performance factor.

The major challenge, though, is faced following the mapping of the reads: data must be turned into biological information. Pivotal for this is the availability of efficient software and strategies for downstream analysis.

In this tutorial you will learn how you can analyze NGS data with the Genomatix system, covering the analysis of RNA-Seq and ChIP-Seq data.

Introduction to Genomatix Genome Analyzer

The Genomatix Genome Analyzer (GGA) is an integrated software/hardware solution for second level analysis of NGS data, after reads have been mapped to the respective genomic target sequences. An easy to use web interface gives access to a broad range of analysis applications for Chip-Seq, RNA-Seq, and DNA-Seq data, among them:

Peak finding
Position data of mapped single reads can be clustered to detect peaks and separate signal from background.

Genome annotation
NGS data can be integrated, correlated, and visualized within the extensive genome annotation in ElDorado. Comparative genomics allows cross-species analysis for phylogenetically conserved regions and regulatory structures.

Expression analysis
The GGA generates normalized transcript expression values from your NGS data and genomic annotation. Compare data sets for differential expression and upload the results into Genomatix Pathway System to generate and analyze gene networks.

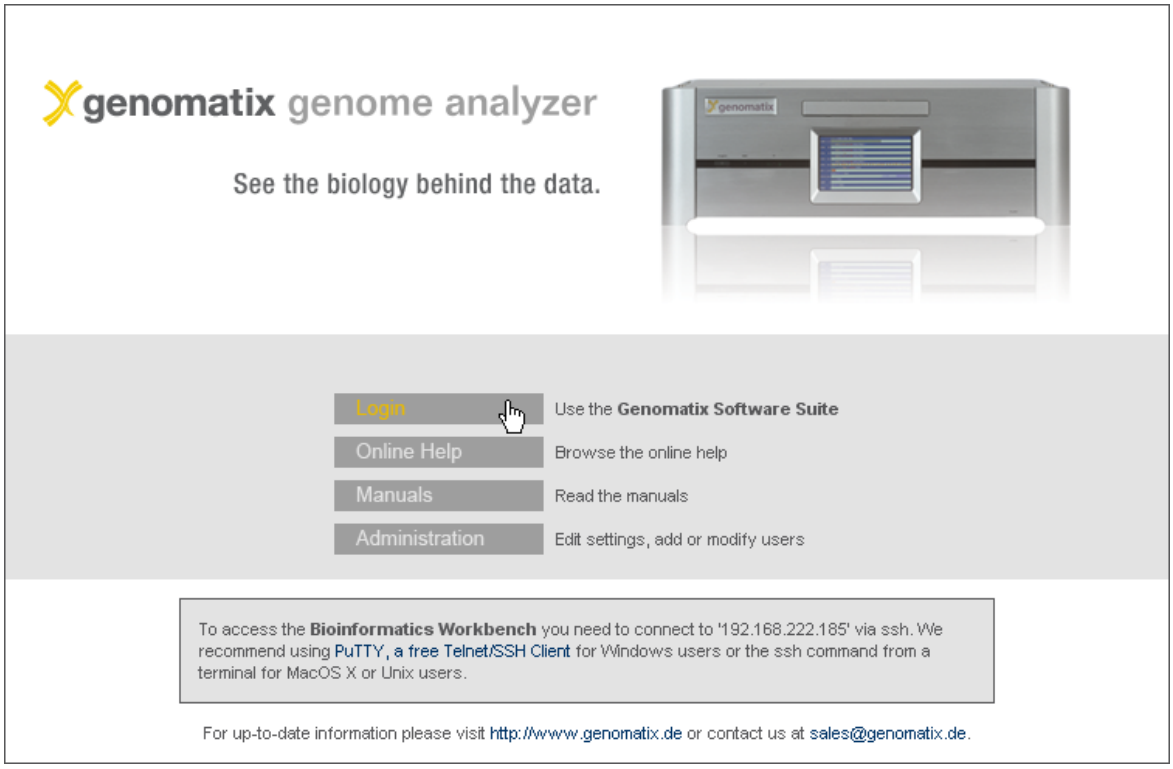
Transcription factor analysis
Genome-wide transcription factor (TF) analysis identifies overrepresented TF binding sites and phylogenetically conserved functional elements. Correlation with genomic annotation finds potential regulatory targets of TF binding. Use CoreSearch for de novo binding site definition from your ChIP-Seq data.

Data meta analysis
Compare several data sets in position correlation graphs, e.g. for the genome wide elucidation of TF interaction, and retrieve regions based on correlation.

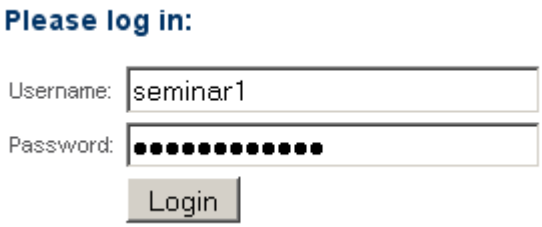
Variant analysis
Genome wide small variant analysis identifies effects on protein sequences and TF binding sites, using the genome and TF binding site annotation in ElDorado and MatBase.

CNV analysis
Pair-wise comparison of BAM files predicting copy number variations, including annotation, filter options, visualization, and links to downstream analysis tools.

Open the home page of the Genomatix Genome Analyzer in your web browser. You should see a page like this:

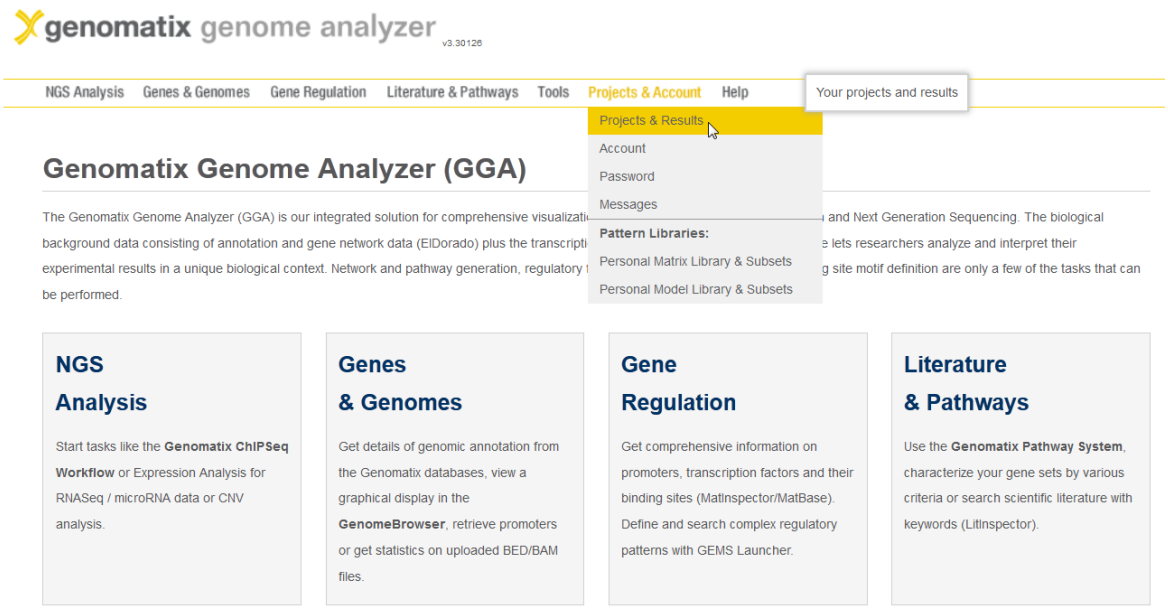


Click the 'Login' button and enter your user name and password:

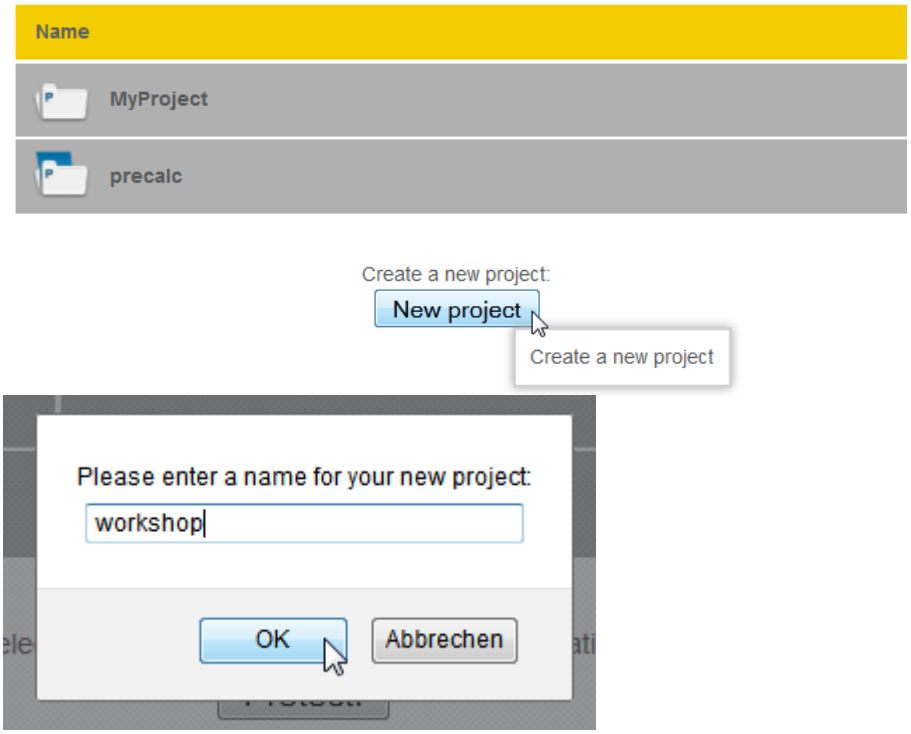


Creating a project

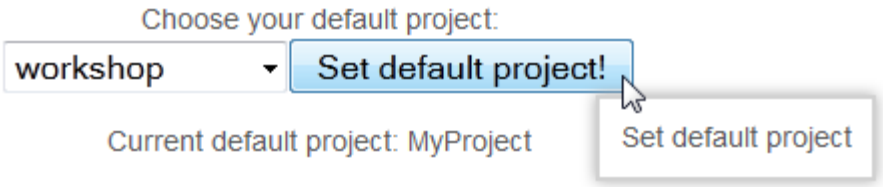
At the top of each page, you'll find a navigation menu bar which allows you to access the available programs. Select the Projects & Results item from the Projects & Account menu.



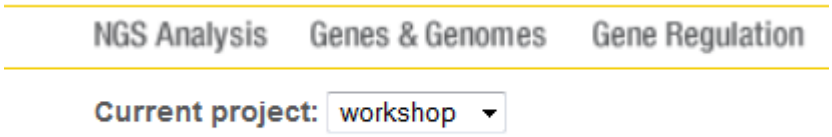
Press the New project button, enter a name for your project in the pop-up dialog, and click on OK.



Using the controls, set the new project as your default project.



The project will be the default in the upper left hand corner project selection on the different program pages.



Data background

Tbx20 transcription factor binding and effects on expression in the adult mouse heart

The following examples are based on publicly available RNA-Seq and ChIP-Seq data from adult mouse heart (accession number GSE30943 on the NCBI Gene Expression Omnibus at http://www.ncbi.nlm.nih.gov/geo).

Tbx20, a transcription factor required for cardiac development, has key roles in early heart development. It has been associated with congenital heart diseases in humans, including defects in septation, chamber growth and valvulogenesis. Conditional ablation of Tbx20 in adult cardiomyocytes leads to a rapid onset and progression of heart failure, with prominent conduction and contractility phenotypes that lead to death. Tbx20 can act both as an activator and a repressor of transcription (Sakabe et al., 2012).

The available data comprise expression data from wild type and Tbx20 knockout adult mouse hearts in triplicates, as well as Tbx20 ChIP-Seq data and input DNA controls from wild type hearts. For this tutorial, sequence files were downloaded from GEO, transferred into fastq format, and mapped to the mouse genome (NCBI build 38) using the Genomatix Mining Station. The genomic positions of the uniquely mapping reads are available in bigBed (*.bb) format on the Genomatix Genome Analyzer server used during the workshop.

RNA-Sequencing analysis

Principle component analysis

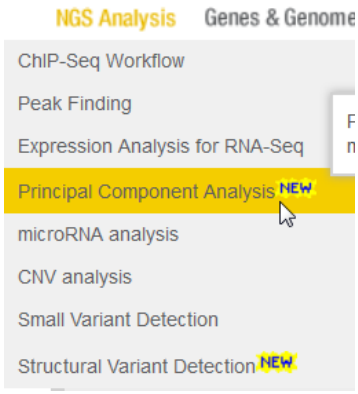
Principal component analysis (PCA) is a statistical procedure that can be used for exploratory data analysis. PCA uses linear combinations of the original data (e.g. gene expression values) to define a new set of unrelated variables (principal components). These new variables are orthogonal to each other, avoiding redundant information.

PCA can be thought of as fitting an n-dimensional ellipsoid to the data, where each axis of the ellipsoid represents a principal component. If some axis of the ellipse is small, then the variance along that axis is also small, and by omitting that axis and its corresponding principal component from our representation of the dataset, we lose only a commensurately small amount of information.

Thus, PCA can be used to reduce the dimensions of a data set, allowing the description of data sets and their variance with a reduced number of variables. Since similarities between data sets are correlated to the distances in the projection of the space defined by the principal components, PCA can also be used to identify outliers with respect to the principal components.

It is often sufficient to look at the first two components, as these describe the largest variability.

A PCA tool can be found in the *NGS Analysis* menu in the navigation bar; please open this now.



This task can be used to get an impression of the similarity of RNA-sequencing samples, i.e. to identify subgroups or outliers.

Based on the read distribution in the input files, a normalized expression value (NE) will be calculated for each locus (or transcript) for each input file. The NE value is based on the number of reads located in the exons of the locus/transcript and is normalized to the length of the locus/transcript and the density of the data set. The resulting NE matrix is then used as input for the PCA, using the R package pcaMethods (Stacklies et al., 2007).

For this analysis, we'll need read position files in BED file format, or as bigBed, the corresponding binary format, or, alternatively, as BAM file.

Here is an example for a BED file:

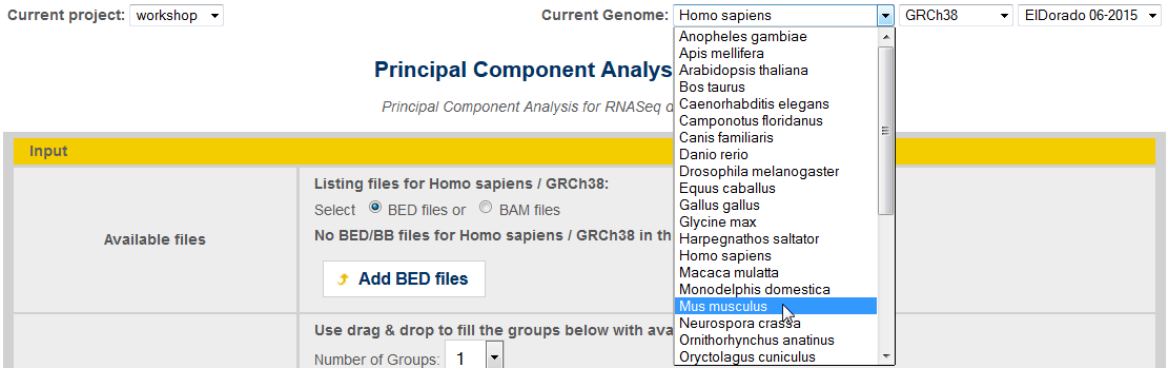
```
chr1  3007329    3007356    4_112_715_245    0.962963    +
chr1  3007329    3007356    4_97_641_338     0.962963    +
chr1  3011584    3011611    4_74_929_759     1.000000    -
chr1  3014985    3015012    4_139_94_580     1.000000    +
chr1  3020759    3020786    4_99_752_96      1.000000    +
chr1  3020873    3020900    4_137_571_605    1.000000    -
chr1  3024593    3024620    4_197_207_931    0.925926    +
chr1  3025020    3025047    4_124_676_441    1.000000    +
chr1  3025020    3025047    4_54_459_727     0.925926    +
chr1  3025914    3025941    4_110_349_304    1.000000    +
chr1  3026179    3026206    4_95_762_768     0.925926    -
chr1  3038718    3038745    4_182_675_953    0.962963    -
```

The first three columns are mandatory:

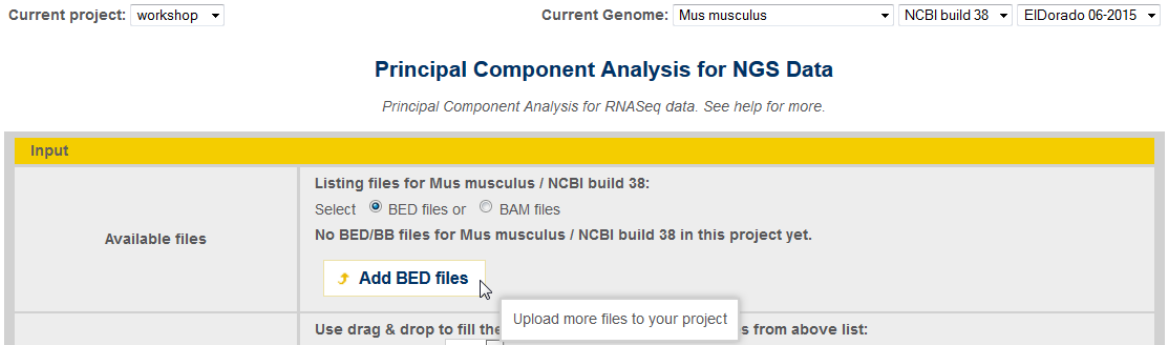
Col 1 : chromosome (starting with chr)
Col 2 : start position of the read (counting starts from 0)
Col 3 : end position of the read (start < end, represents the last nucleotide of the sequence + 1)

Additional optional information can be provided in the next columns; it is important that the order of the columns is maintained, i.e. if the file contains strand information, it must be placed in column 6, and both columns 4 and 5 cannot be empty.
Col 4 : SeqId (alpha-numerical value, <=50 characters)
Col 5 : Score (usually the quality score of the mapping)
Col 6 : strand information
    + : plus strand
    - : minus strand
    0 : no strand information available

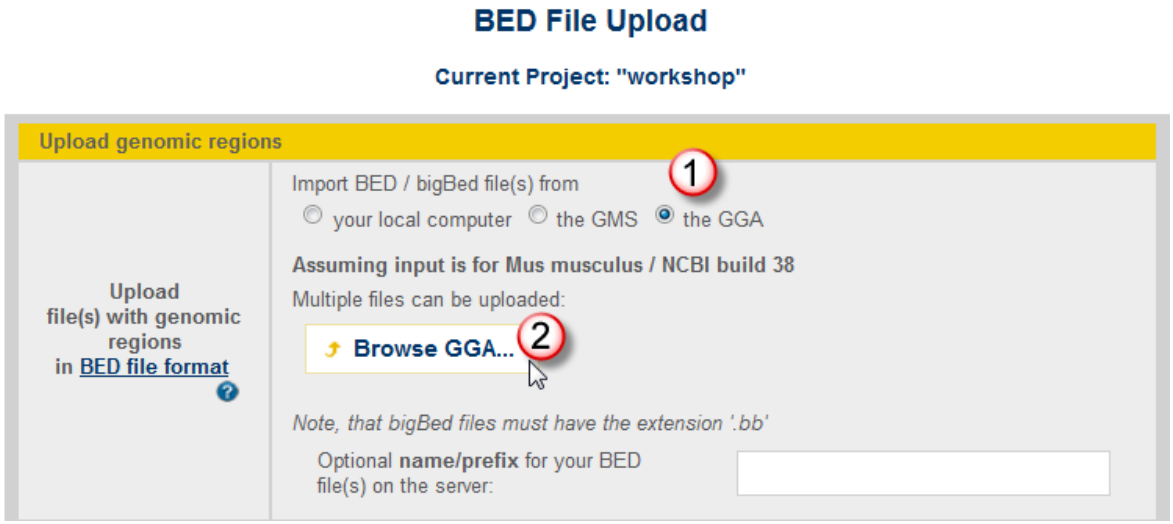
As you will work with mouse data, use the controls in the upper right hand corner of the input page to change the current genome selection to *Mus musculus*.



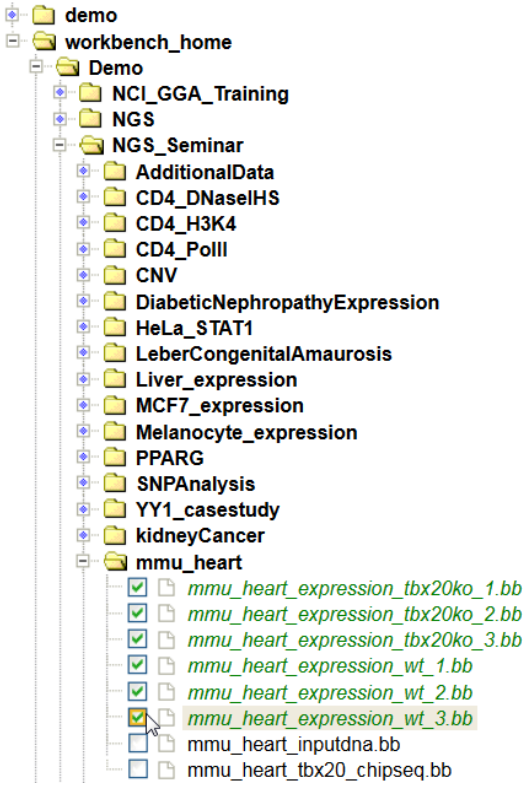
Press the *Add BED files* button to open a dialog for adding BED or bigBed files to your project.



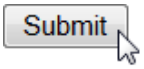
Then select *Import from the GGA*, and press the *Browse GGA* button.



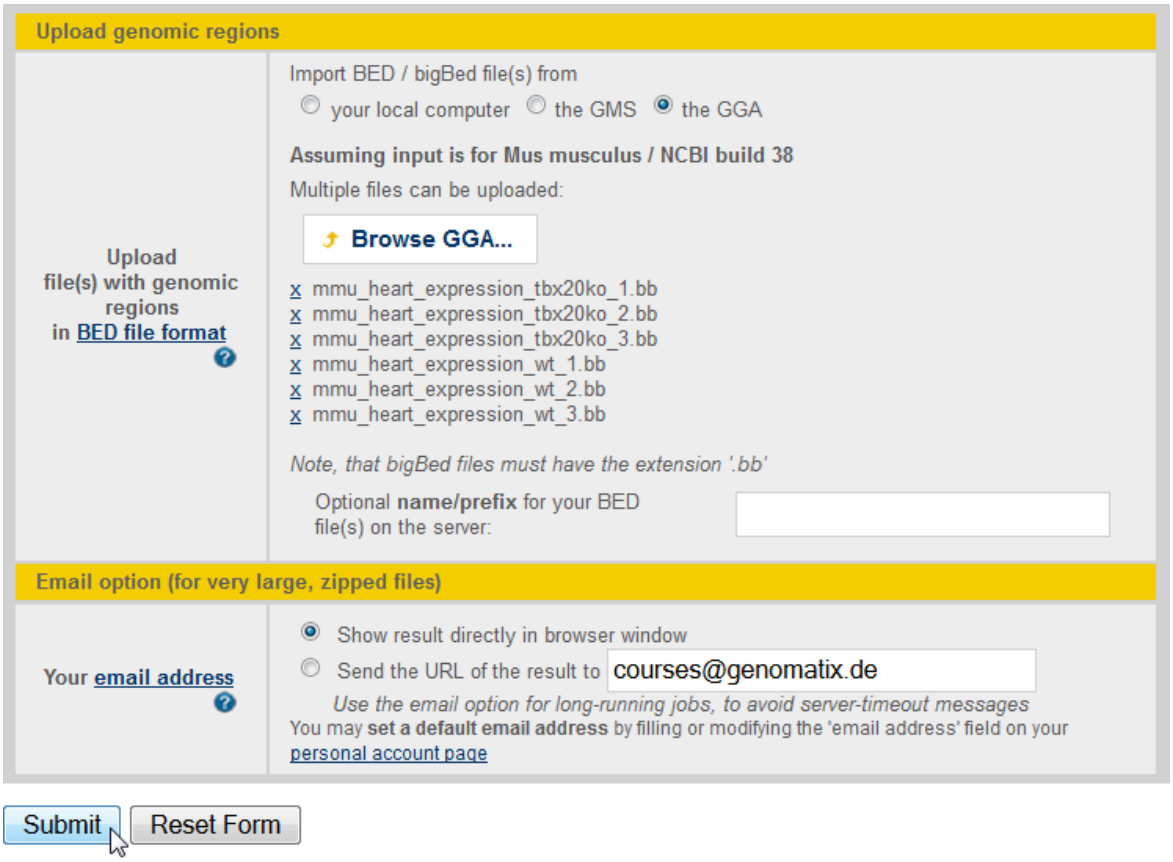
Open the directory structure until you come to the subdirectory at the path *workbench_home/Demo/NGS_Seminar/mmu_heart*. There you'll find the files with the expression data, and also the ChIP-Seq files which we will use later. For now, select the first 6 files starting with *mmu_heart_expression* by ticking the check boxes.



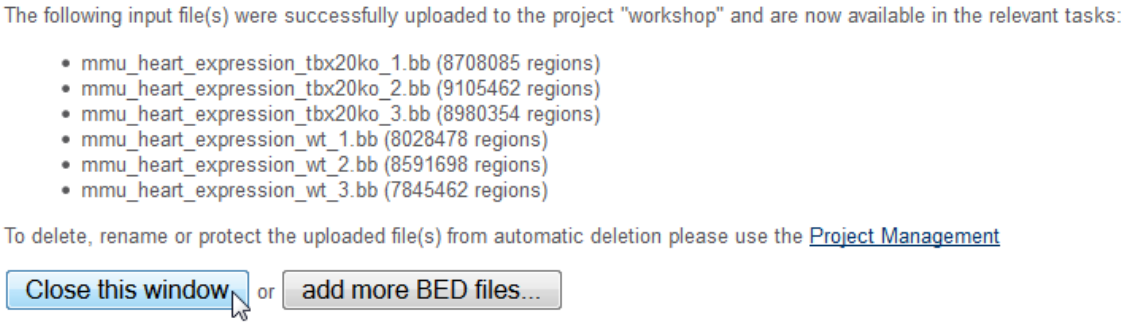
Press the Submit button at the bottom of the file selection dialog to close it.



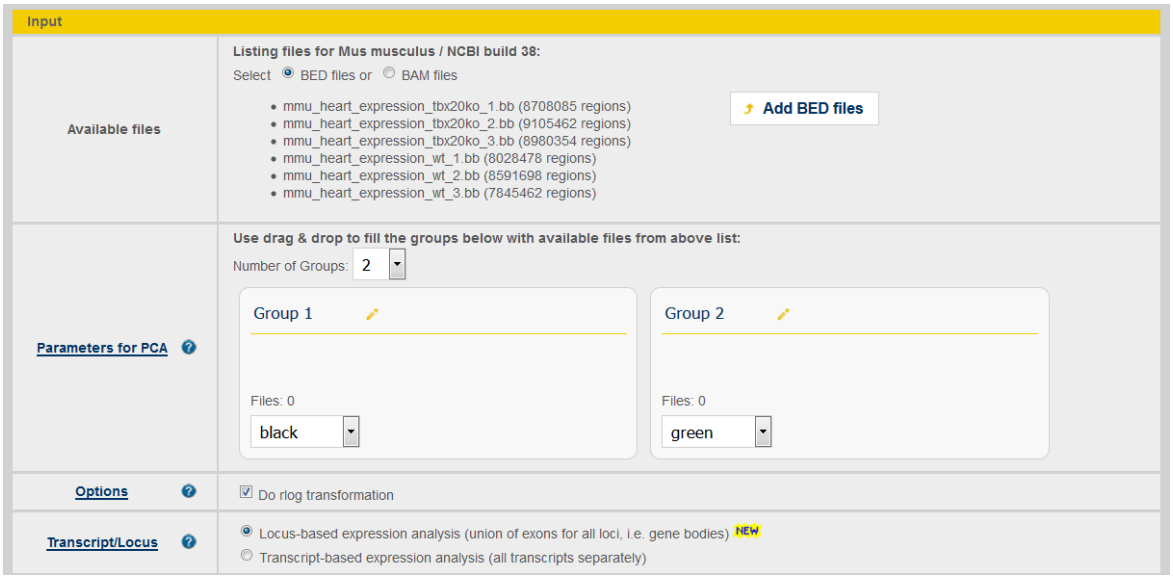
In the upload dialog, press *Submit*.



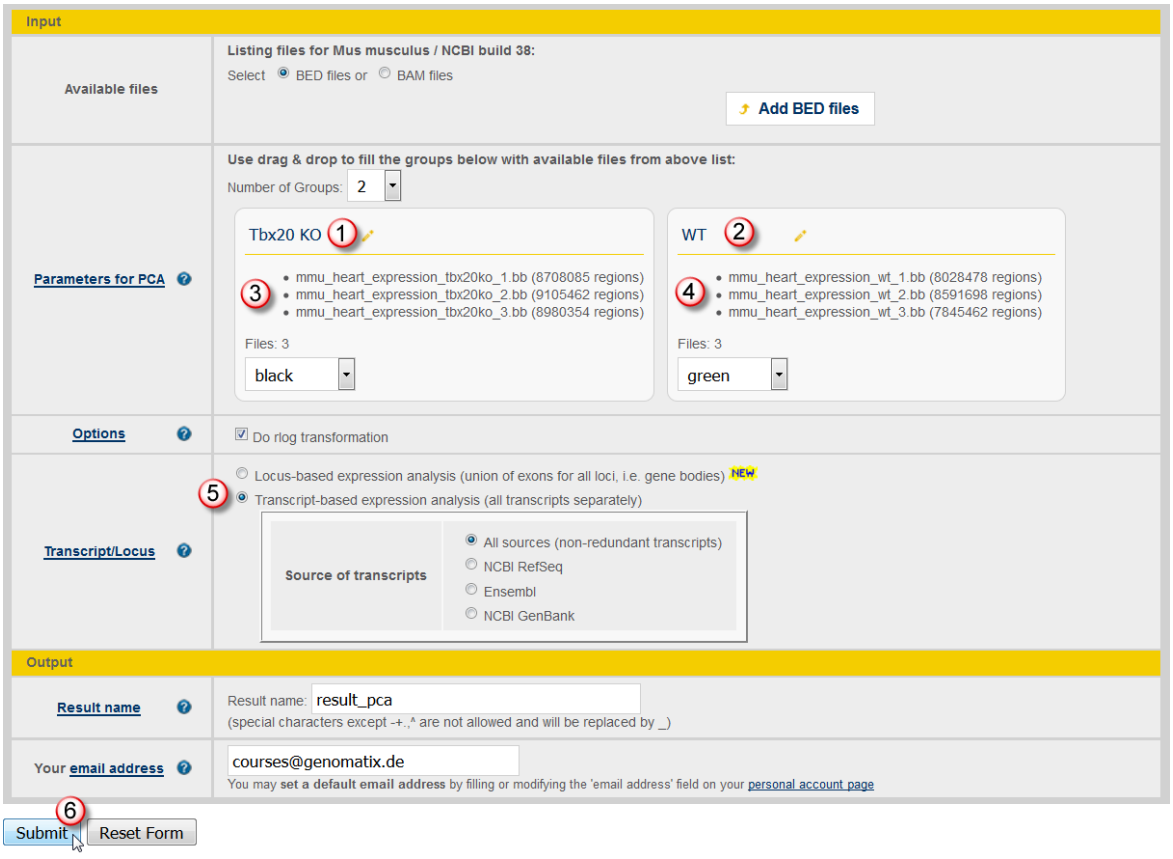
The upload will start; when it is finished, press the *Close this window* button in the dialogue.



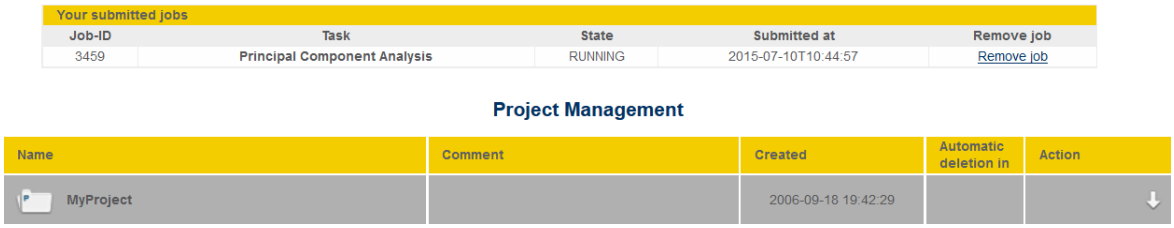
The uploaded files will be listed as below.



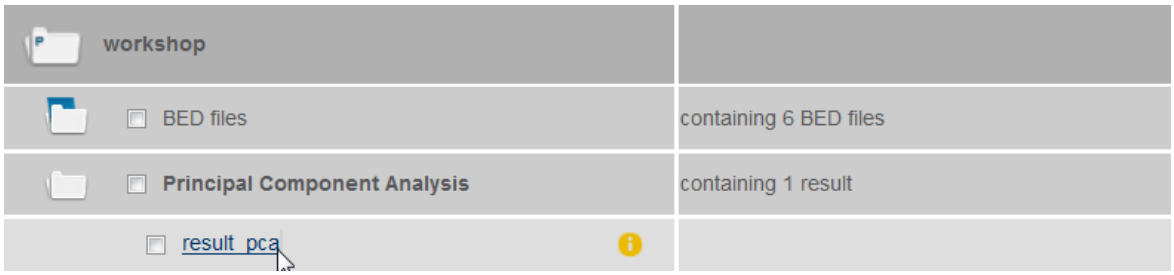
Rename the groups, e.g. Group 1 to *Tbx20 KO*, Group 2 to *WT*. Drag & drop the files into the corresponding group fields. Select the transcript-based analysis (for consistency with the comparative expression analysis that we'll run later) and submit the job, which will run in the background.



Check the *Project Management* page to see running jobs. The PCA analysis will be listed as *RUNNING* or *PENDING* (in case it's waiting for a free processor core). Please note that the list does not automatically update; if you wish to see the current state, reload the page.

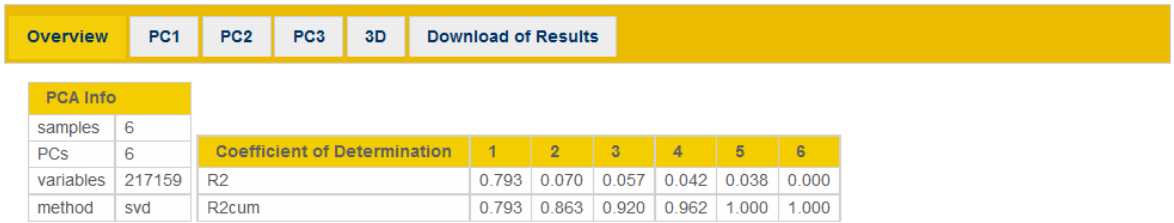


When he job is finished, the result will appear in the current project under *Principal Component Analysis*. Click on the result name to display the result.



The Overview page displays the overview table and a number of analytic plots.

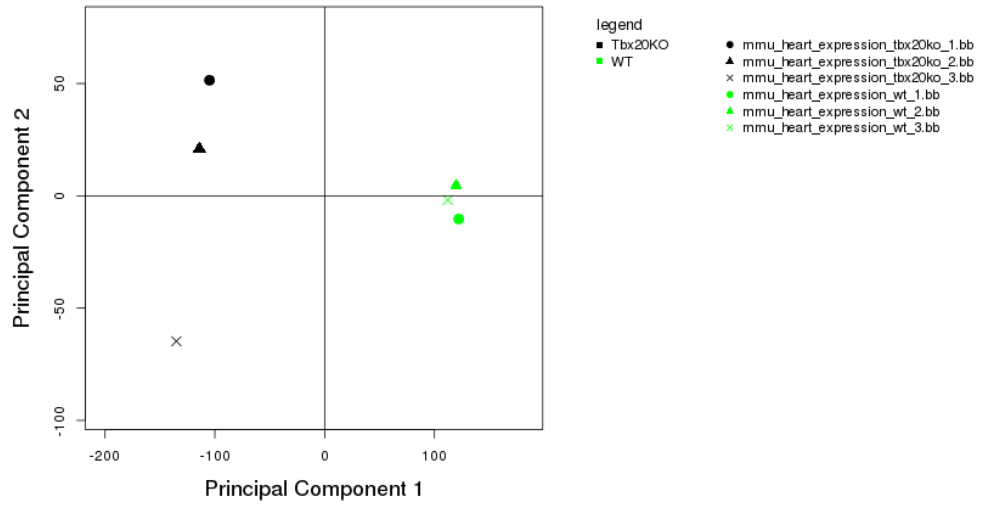
Samples	Number of samples submitted to analysis
PCs	Number of principal components calculated (max 10)
Variables	Number of loci or transcripts considered for analysis
Method	svd = singular value decomposition
R2	The proportion of variance explained by each PC calculated (eigenvalue)
R2cum	The cumulative proportion of the variance explained by the current and all preceding principal components.



Score plot

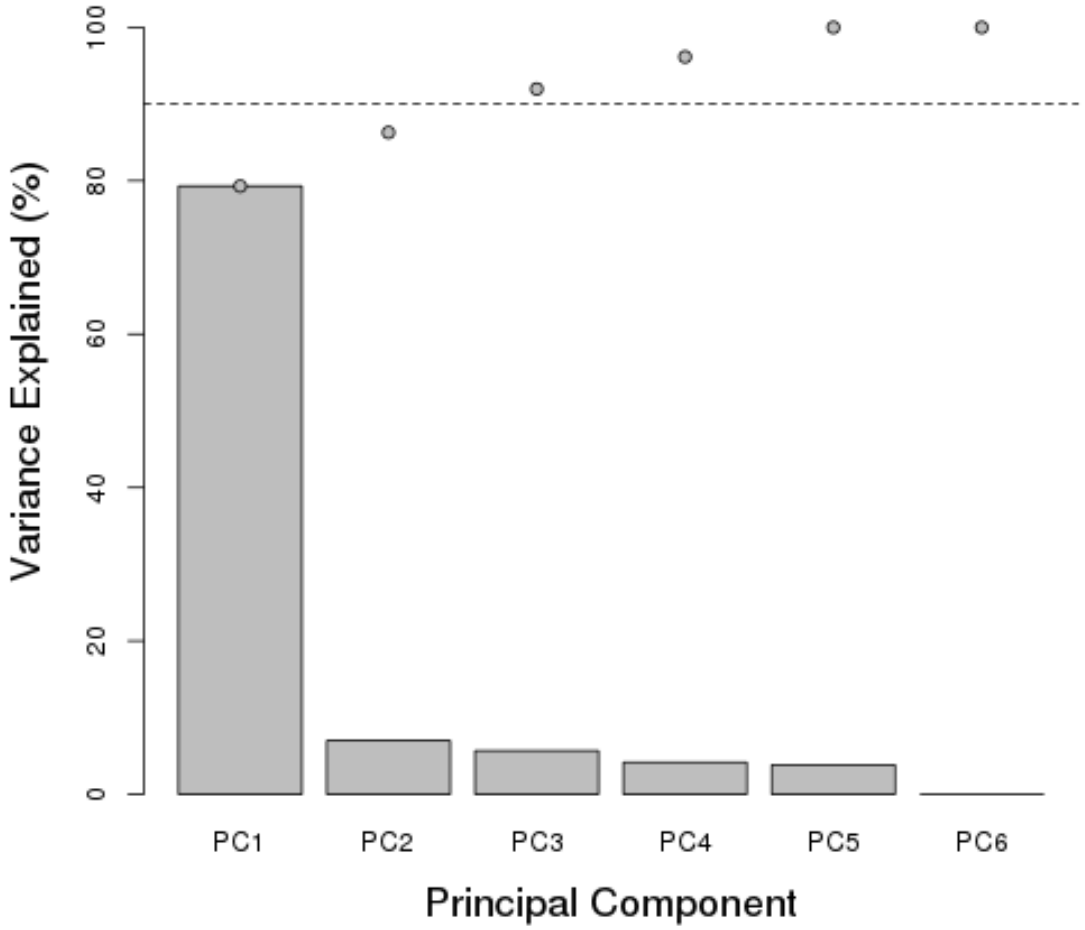
The score plot displays each sample in the data set with respect to the first two principal components and can therefore be used to interpret the relations among the samples. This information can be used to identify outliers.

In this data set, replicates from the WT group show high similarity with respect to the first two principal components. Replicates in the Tbx20 KO group show a greater variation, mainly due to the values for replicate 3. However, the two groups separate from each other.



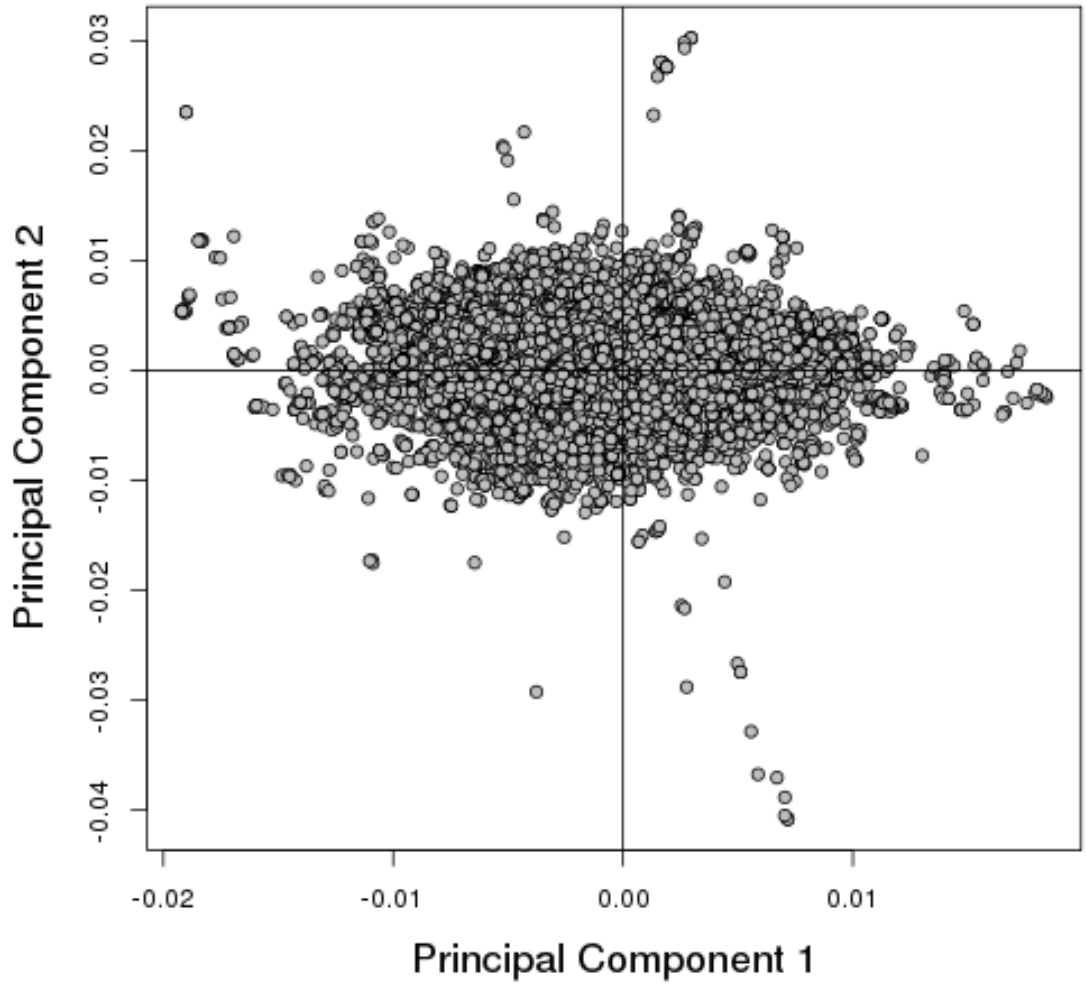
Scree plot

The scree plot visualizes which principal components account for which fraction of total variance in the data. The principal components are listed by decreasing order of contribution to the total variance. The bars show the proportion of variance represented by each component (R2) and the points shows the cumulative variance (R2cum). In this case, the first component explains almost 80% of the total variance, the first three components together over 90% of it.



Loadings plot

The loadings plot is a plot of the relationship between original variables (genes) and subspace dimensions. It summarizes correlation and anti-correlation of genes/transcripts with the first two principal components.



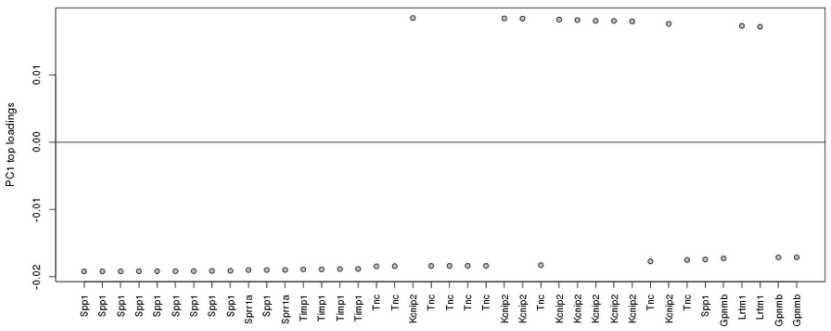
Details for principal components

For the top principal components that are needed to account for 90% of the variance in the data (or up to a maximum of 10 PCs) the 40 transcripts/loci with the highest absolute loadings are shown in a table and a plot.

In the current example, the first 3 PCs account for >90% of the variance; below you see part of the results for the first component. Please note that a gene name can be listed several times for transcript-based analyses.

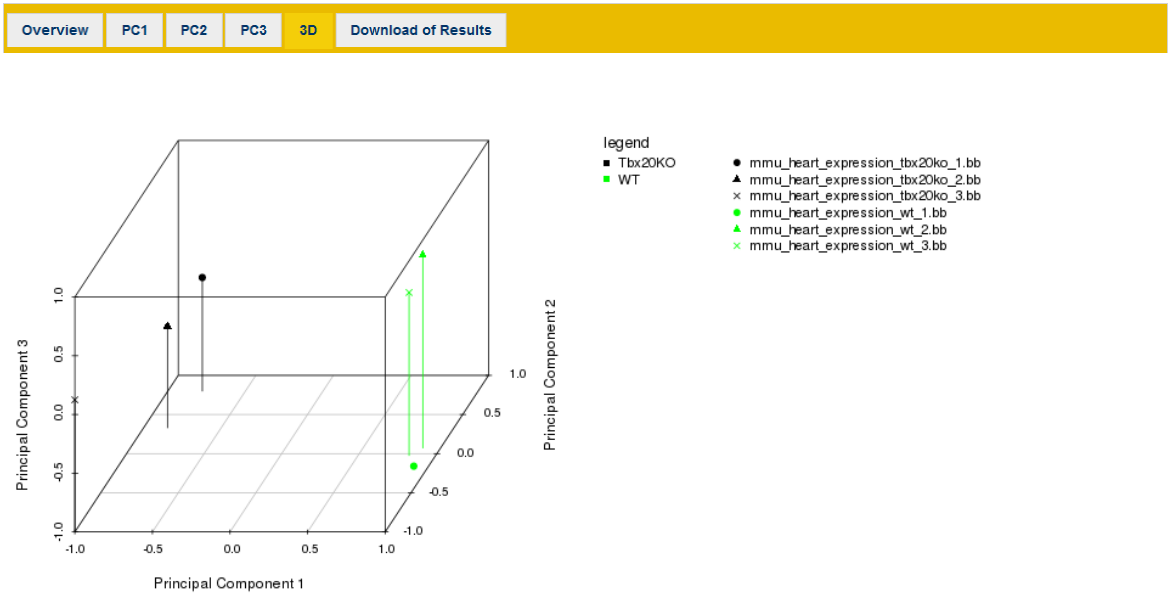
Top 40 Loadings for PC1

Rank	GeneID	Symbol	Loading
1	20750	Spp1	-0.0192
2	20750	Spp1	-0.0192
3	20750	Spp1	-0.0192
4	20750	Spp1	-0.0192
5	20750	Spp1	-0.0192
6	20750	Spp1	-0.0192
7	20750	Spp1	-0.0192
8	20750	Spp1	-0.0191
9	20750	Spp1	-0.0191
10	20753	Sprr1a	-0.0190
11	20750	Spp1	-0.0190
12	20753	Sprr1a	-0.0190
13	21857	Timp1	-0.0189
14	21857	Timp1	-0.0189
15	21857	Timp1	-0.0189
16	21857	Timp1	-0.0189
17	21923	Tnc	-0.0185
18	21923	Tnc	-0.0184



3D score plot

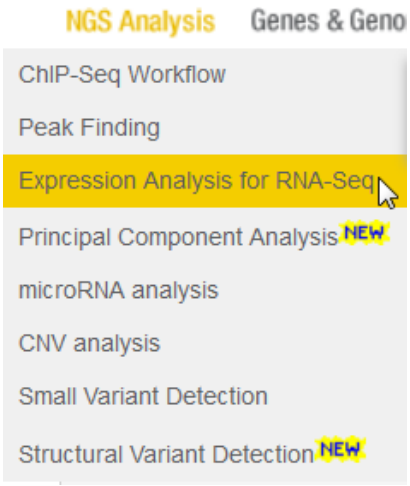
This score plot displays each sample in the data set with respect to the first three principal components.



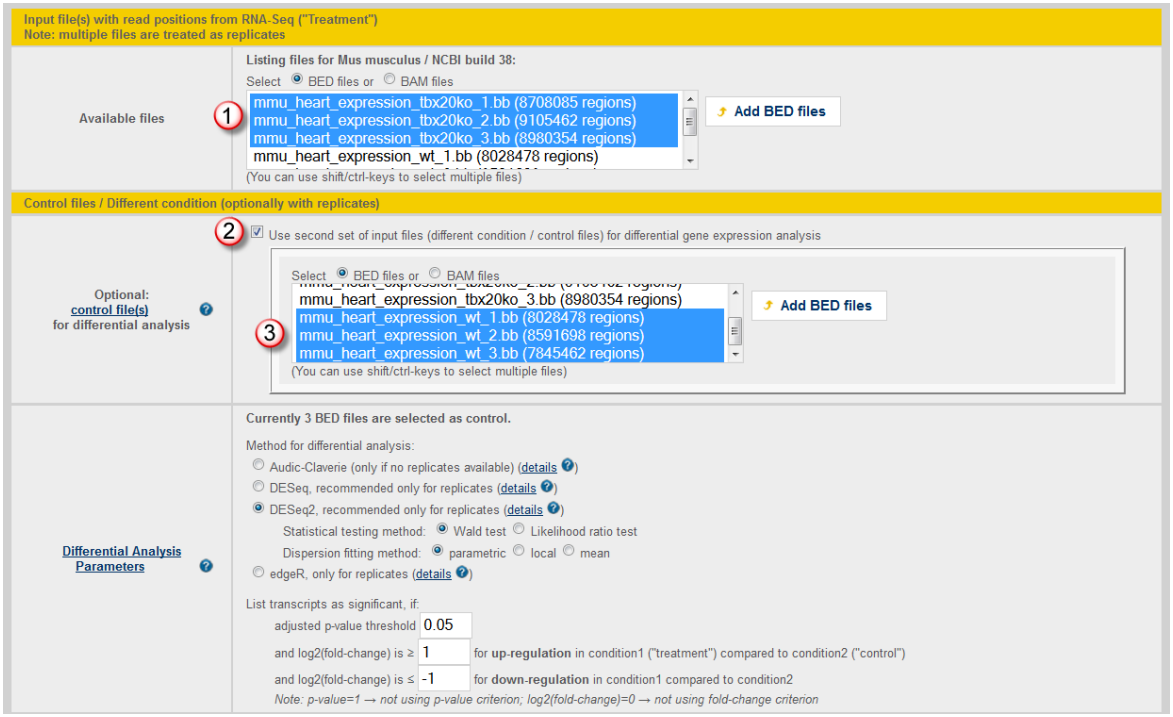
Differential expression in Tbx20-/- knockout compared to wild type adult mouse hearts

Comparative expression analysis

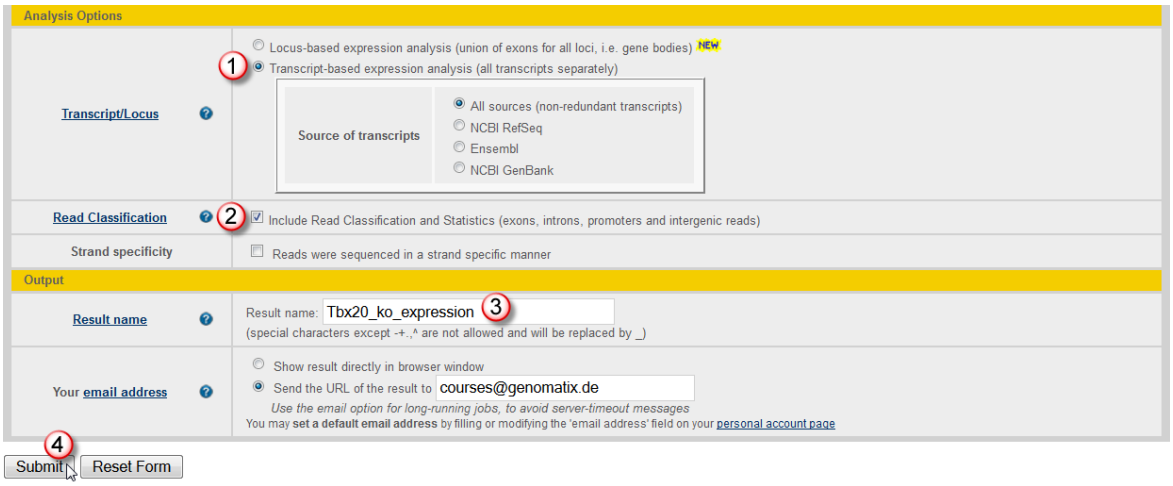
In this example, we'll carry out a differential expression analysis with the files that were subjected to a PCA in the previous step. Open the input page for *Expression Analysis for RNA-Seq* from the *NGS Analysis* menu:



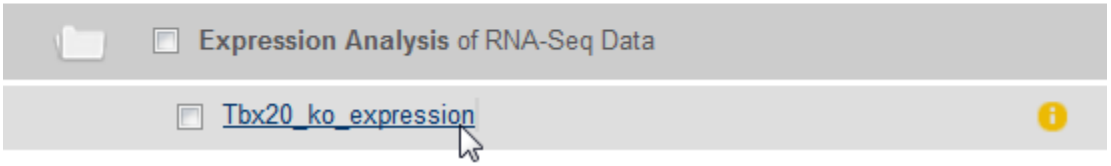
Select the *tbx20ko* files to use them as the treatment group. Then tick the *Use second set...* checkbox and select the *wt* files in the second list as controls. You can choose from a number of methods and parameter settings for differential expression analysis. For this example, please leave the default settings for analyses with replicates: DESeq2 using the Wald test with parametric dispersion fitting.



Note that you have the option to run the analysis locus-based or transcript based. For this example, please take the *transcript-based* option. In this case, you can then choose from different transcript annotations. Please leave the latter at the default, activate the *read classification*, provide a result name, and run the analysis in the background, which should take about 10 minutes.



After completion, load the result from the project management page.



Different files with analysis results on transcript and gene level can be downloaded. Of 217159 annotated transcripts, 31049 are differentially expressed (17021 up-regulated, 14028 down-regulated), corresponding to 4927 genes (2729 up-regulated, 2214 down-regulated).

Differential Expression Overview

	Transcripts	Genes (known GeneId)
Total number analyzed	217159 download details (tab-separated) (62Mb)	29812 download details (tab-separated) (2.2Mb)
Differential expression	31049 download details (tab-separated) (9.0Mb)	4927
Up-regulation	17021 Download BED file of Transcripts (1.1Mb) Save BED file to project management	2729 download details (tab-separated) (288Kb) download gene list (52Kb)
Down-regulation	14028 Download BED file of Transcripts (881Kb) Save BED file to project management	2214 download details (tab-separated) (244Kb) download gene list (48Kb)
Up- and down-regulated genes (with different transcripts)	-	16 download details (tab-separated) (4.0Kb)

Click the *download details* link for the differentially expressed transcripts, and open the file in a spreadsheet program; this will show you the list of the transcripts which are regulated according to the selected analysis method and thresholds (adjusted p-value ≤ 0.05; log2 fold change ≥ 1 or ≤ -1) including detailed information. NE (normalized expression) and RPKM (reads per thousand base pairs per million mapped reads) values are used as measures for expression. The output below is broken down into three blocks.

TranscriptId	Accn	LocusId	Symbol	GeneId	ContigAccn	Chromosome	Strand	Start	End	Length	#exons	p-value	adj. p-value	log2(fold change)	Regulation
GXT_12942264	AK090041	GXL_1787596	Slamf9	98365	NC_000067	chr1	+	172475374	172478575	1297	4	2.55E-04	1.06E-03	1.05	up
GXT_12942270	AK089400	GXL_1196246	Kif21b	16565	NC_000067	chr1	+	136131454	136149993	2606	6	2.24E-05	1.14E-04	2.16	up
GXT_12942315	AK088077	GXL_87684	Trmt1l	98685	NC_000067	chr1	+	151428666	151436707	1528	3	1.25E-03	4.42E-03	-1.17	down
GXT_12942316	AK088027	GXL_742666	Cd48	12506	NC_000067	chr1	+	171682009	171705256	920	3	1.26E-02	3.36E-02	1.62	up
GXT_12942320	AK087631	GXL_20287	Irf6	54139	NC_000067	chr1	+	193153154	193166868	1760	5	2.18E-03	7.21E-03	1.85	up
GXT_12942323	AK087427	GXL_87676	Arpc5	67771	NC_000067	chr1	+	152766676	152775503	1687	4	8.91E-14	1.33E-12	1.41	up
GXT_12942325	AK086974	GXL_110155	Stradb	227154	NC_000067	chr1	+	58973641	58991512	1229	7	4.23E-07	2.88E-06	-1.15	down
GXT_12942344	AK085015	GXL_6599	Arid5a	214855	NC_000067	chr1	+	36307760	36322975	4444	5	1.67E-13	2.42E-12	1.6	up
GXT_12942346	AK084971	GXL_110144	Fastkd2	75619	NC_000067	chr1	+	63730651	63753385	3179	12	3.20E-06	1.89E-05	-1.02	down
GXT_12942350	AK084836	GXL_110247	AK084836	0	NC_000067	chr1	+	74295592	74297905	1745	2	7.56E-03	2.15E-02	1.35	up

...

#reads treat1	#reads treat2	#reads treat3	#reads ctrl1	#reads ctrl2	#reads ctrl3	NE treat1	NE treat2	NE treat3	NE ctrl1	NE ctrl2	NE ctrl3	mean NE(treat)	stddev NE(treat)	mean NE(ctrl)	stddev NE(ctrl)
102	79	79	24	31	33	0.09079	0.06781	0.0689	0.02337	0.02824	0.03292	0.07583	0.01059	0.02818	0.0039
29	32	33	5	5	3	0.01253	0.01282	0.01259	0.00242	0.00208	0.00149	0.01265	0.00012	0.002	0.00038
39	28	12	45	50	38	0.02899	0.02002	0.00888	0.0364	0.03834	0.03185	0.0193	0.00823	0.03553	0.00272
35	12	10	3	8	0	0.04435	0.01452	0.0123	0.00412	0.01027	0	0.02372	0.01461	0.0048	0.00422
20	19	16	2	5	2	0.01259	0.01075	0.01028	0.00072	0.00336	0.00147	0.01121	0.001	0.00185	0.00111
414	395	577	144	116	109	0.28403	0.26067	0.38556	0.10707	0.08124	0.0813	0.31009	0.05421	0.08987	0.01216
76	73	59	99	120	116	0.0645	0.05979	0.0497	0.0925	0.09902	0.10633	0.058	0.00617	0.09928	0.00565
286	447	503	111	81	94	0.07477	0.11098	0.12669	0.03127	0.02154	0.02737	0.10415	0.02174	0.02673	0.004
97	88	68	127	131	114	0.03447	0.03082	0.02313	0.05007	0.04683	0.04558	0.02947	0.00473	0.04749	0.00189
13	30	43	7	6	9	0.00869	0.01914	0.02787	0.00507	0.00406	0.00667	0.01857	0.00784	0.00527	0.00107

...

RPKM treat1	RPKM treat2	RPKM treat3	RPKM ctrl1	RPKM ctrl2	RPKM ctrl3	mean RPKM(treat)	stddev RPKM(treat)	mean RPKM(ctrl)	stddev RPKM(ctrl)
9.03104	6.68937	6.78256	2.30483	2.78191	3.24306	7.50099	1.08258	2.7766	0.38305
1.27791	1.34857	1.41009	0.23898	0.22331	0.14673	1.34552	0.05401	0.20301	0.0403
2.93102	2.01249	0.87451	3.66822	3.80862	3.16987	1.93934	0.84116	3.5489	0.27408
4.36875	1.43249	1.21037	0.40616	1.0121	0	2.3372	1.43938	0.47275	0.41586
1.30495	1.1856	1.01231	0.14154	0.33066	0.14484	1.16762	0.12014	0.20568	0.08838
28.1814	25.71461	38.08617	10.63198	8.0032	8.23556	30.66073	5.34629	8.95691	1.18824
7.10132	6.52332	5.34573	10.03345	11.3645	12.03061	6.32346	0.73052	11.14285	0.83026
7.39042	11.04667	12.60377	3.11111	2.12145	2.6961	10.34695	2.18509	2.64289	0.40578
3.50396	3.04012	2.38191	4.976	4.79625	4.57084	2.97533	0.46036	4.78103	0.16576
0.85551	1.8881	2.74397	0.49965	0.4002	0.6574	1.82919	0.77208	0.51908	0.1059

An unfiltered file with the same structure listing all analyzed transcripts is also available.

For detailed result lists on gene level, click on the corresponding links in the rightmost column of the differential expression overview. For example, the top of the list of down-regulated genes looks like this:

GeneId	Symbol	#transcripts regulated	total #transcripts for gene	mean log2(fold change) of reg. trans.	min fold change of reg. trans.	max fold change of reg. trans.	fc stddev	min p_value
80906	Kcnip2	13	13	-6.141	-6.409	-5.799	0.19	2.05E-230
68052	Rps13	3	7	-6.139	-6.139	-6.139	0	6.16E-14
13643	Efnb3	3	3	-5.773	-5.796	-5.76	0.016	1.16E-152
319476	Lrtm1	5	5	-5.391	-5.516	-5.276	0.091	5.28E-123
142687	Asb14	8	8	-5.305	-5.737	-3.437	0.73	7.60E-74
319942	A530016L24Rik	6	6	-5.299	-5.457	-5.125	0.109	7.50E-65
30952	Cngb3	3	3	-5.081	-5.165	-4.946	0.097	2.03E-10
213402	Armc2	10	12	-4.801	-5.495	-2.308	0.877	8.12E-42
78910	Asb15	7	7	-4.791	-5.115	-4.349	0.288	8.58E-57
238564	Mylk4	10	11	-4.64	-4.902	-3.6	0.352	4.09E-114

...

mean NE(treat.reg.)	stddev NE(treat.reg.)	mean NE(ctrl.reg.)	stddev NE(ctrl.reg.)	mean RPKM(treat.reg.)	stddev RPKM(treat.reg.)	mean RPKM(ctrl.reg.)	stddev RPKM(ctrl.reg.)
0.01348	0.005	0.81806	0.229	1.32801	0.506	81.58881	22.09
0	0	0.28783	0.165	0	0	28.36193	16.262
0.00722	0.001	0.32802	0.043	0.71178	0.106	32.45018	4.281
0.01947	0.01	0.72033	0.236	1.93015	0.94	71.11432	23.38
0.00445	0.002	0.18749	0.06	0.46146	0.266	19.15481	5.898
0.00459	0.001	0.16635	0.039	0.46903	0.13	16.40701	3.813
0	0	0.00876	0.003	0	0	0.88762	0.325
0.00179	0.002	0.0375	0.008	0.17655	0.219	3.77305	0.785
0.00683	0.004	0.19036	0.042	0.67324	0.381	18.99107	4.015
0.00964	0.004	0.21467	0.083	0.96041	0.442	21.88284	8.548

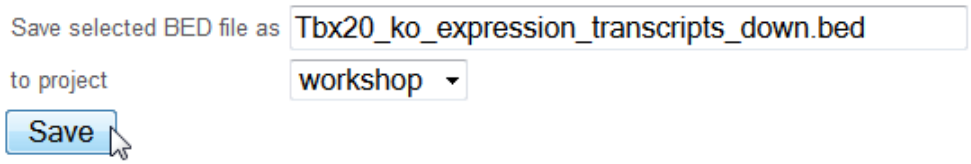
You can download a simple list of regulated genes with Gene IDs, log2 fold changes, and gene symbols.

80906	-6.141	Kcnip2
68052	-6.139	Rps13
13643	-5.773	Efnb3
319476	-5.391	Lrtm1
142687	-5.305	Asb14
319942	-5.299	A530016L24Rik
30952	-5.081	Cngb3
213402	-4.801	Armc2
78910	-4.791	Asb15
238564	-4.64	Mylk4

For later comparison with the Tbx ChIP-Seq data, we'll use the BED file with the positions of the down-regulated transcript. Please save this now to your project management. Click the *Save BED file* link for the down-regulated transcripts.

Up-regulation	17021 Download BED file of Transcripts (1.1Mb) Save BED file to project management	2729 download details (tab-separated) (288Kb) download gene list (52Kb)
Down-regulation	14028 Download BED file of Transcripts (881Kb) Save BED file to project management	2214 download details (tab-separated) (244Kb) download gene list (48Kb)
Up- and down-regulated genes (with different transcripts)	-	16 download details (tab-separated) (4.0Kb)

On the next page, provide a name for the BED file and press the *Save* button.



Next, please download the gene lists of the up-regulated and of the down-regulated genes to your local computer; we will use them later.

Differential Expression Overview

	Transcripts	Genes (known GeneId)
Total number analyzed	217159 download details (tab-separated) (62Mb)	29812 download details (tab-separated) (2.2Mb)
Differential expression	31049 download details (tab-separated) (9.0Mb)	4927
Up-regulation	17021 Download BED file of Transcripts (1.1Mb) Save BED file to project management	2729 download details (tab-separated) (288Kb) download gene list (52Kb) 1
Down-regulation	14028 Download BED file of Transcripts (881Kb) Save BED file to project management	2214 download details (tab-separated) (244Kb) download gene list (48Kb) 2
Up- and down-regulated genes (with different transcripts)	-	16 download details (tab-separated) (4.0Kb)

After you've saved the files, please go back to the output page. The top 5 and top 50 up- and down-regulated genes are also available on the HTML page:

Up-Regulation:

Genes with the highest log2(fold change) for up-regulated Transcripts
in input file(s) (mmu_heart_expression_tbx20ko_1.bb, ...) compared to control file(s) (mmu_heart_expression_wt_1.bb, ...):

Symbol	GeneId	mean log2(fold change) of up-reg. transcripts
Spp1	20750	7.01
Timp1	21857	6.54
Sprr1a	20753	5.91
Bglap3	12095	5.90
Tnc	21923	5.85
>>> show more genes <<< (top 50)		

Down-Regulation:

Genes with the smallest log2(fold change) for down-regulated Transcripts
in input file(s) (mmu_heart_expression_tbx20ko_1.bb, ...) compared to control file(s) (mmu_heart_expression_wt_1.bb, ...):

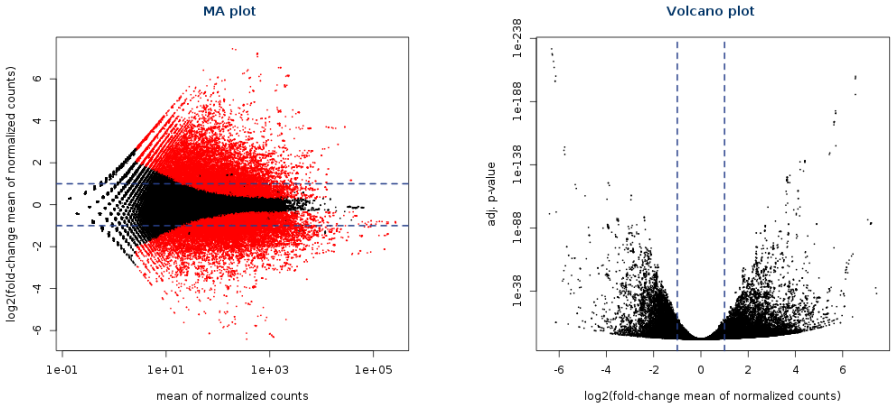
Symbol	GeneId	mean log2(fold change) of down-reg. transcripts
Kcnip2	80906	-6.14
Rps13	68052	-6.14
Efnb3	13643	-5.77
Lrtm1	319476	-5.39
Asb14	142687	-5.30
>>> show more genes <<< (top 50)		

The top up- and down-regulated genes can directly be used as input for the Genomatix Pathway System from the result page (see next step).

Four different diagnostic plots can be viewed and downloaded. The first two are an MA plot and a volcano plot. Points represent transcripts, dashed lines are fold change thresholds.

Left: MA plot (log2 fold-change mean of normalized counts (y-axis) vs. mean of normalized counts (x-axis)). Red dots represent values for significantly regulated transcripts (according to the adjusted p-value, but not taking the log2 fold-change into account). Note that no transcripts with a mean below ~10 normalized counts are considered regulated

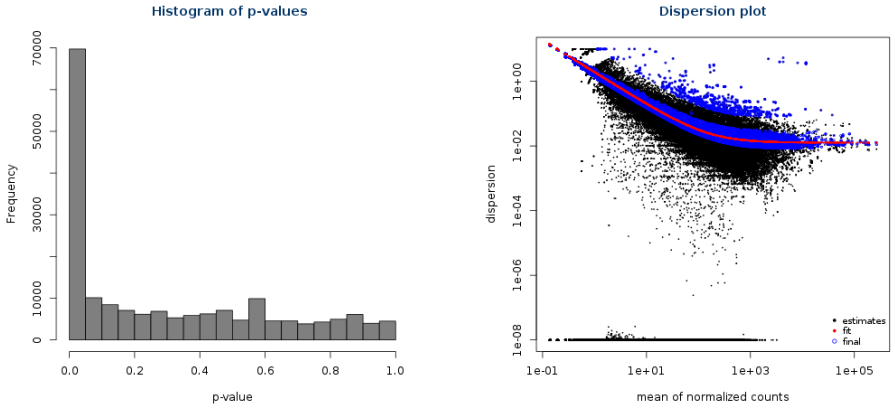
Right: volcano plot of adjusted p-value (y-axis, inverted scale) vs. log2 fold-change mean of normalized counts (x-axis). The volcano plot shows statistical significance (p-value) and biological significance (effect size as log2 fold change) in one graph.



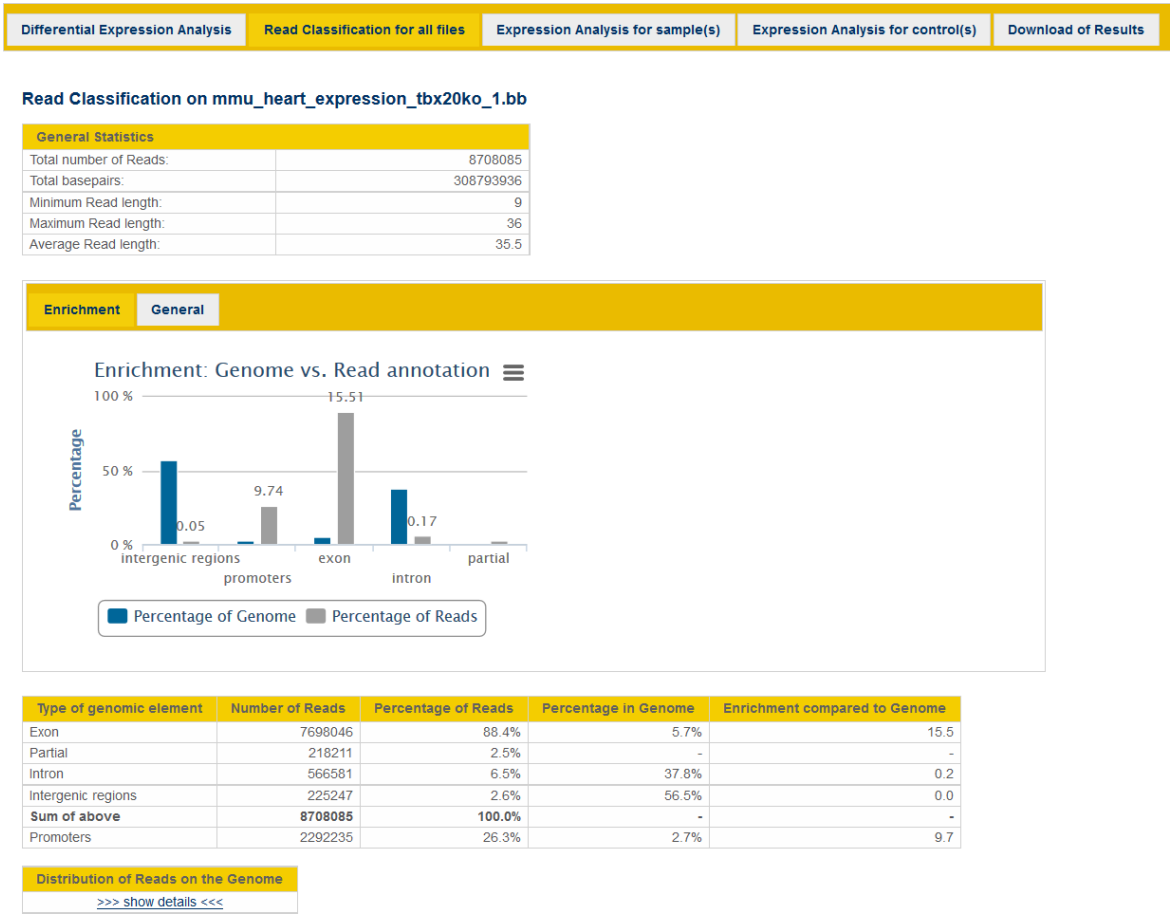
The next two are p-value histogram and a dispersion plot.

Left: p-value histogram showing the distribution of observed p-values in bins of 0.05. As expected for a comparison with significant differences, there is an enrichment of small p-values.

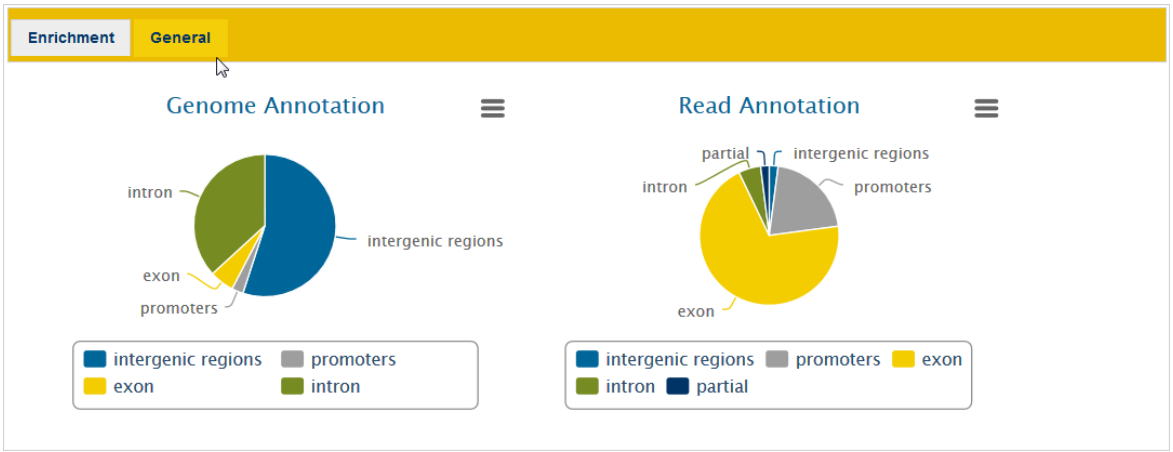
Right: dispersion plot. The dispersion quantifies the within-group variability of each transcript. Black dots: transcript-wise dispersion estimates. Red line: trend line showing the dispersions' dependence on the mean; its shape is influenced by the selected dispersion fitting method. Blue dots near the trend line: final (shrunk towards the trend line) dispersion estimates. Blue dots above main cloud: dispersion outliers, which are not shrunk towards the trend line. Values represented by blue dots are used for significance testing.



The next part shows the read classification for all input files. It also provides enrichment graphs; below are the numbers for one of the knockout samples:



The read classification results can also be shown as pie charts; the left graph shows the fractions of the different annotations in the genome; the right diagram shows the percentages of the corresponding read annotations:



Biology of differentially expressed genes

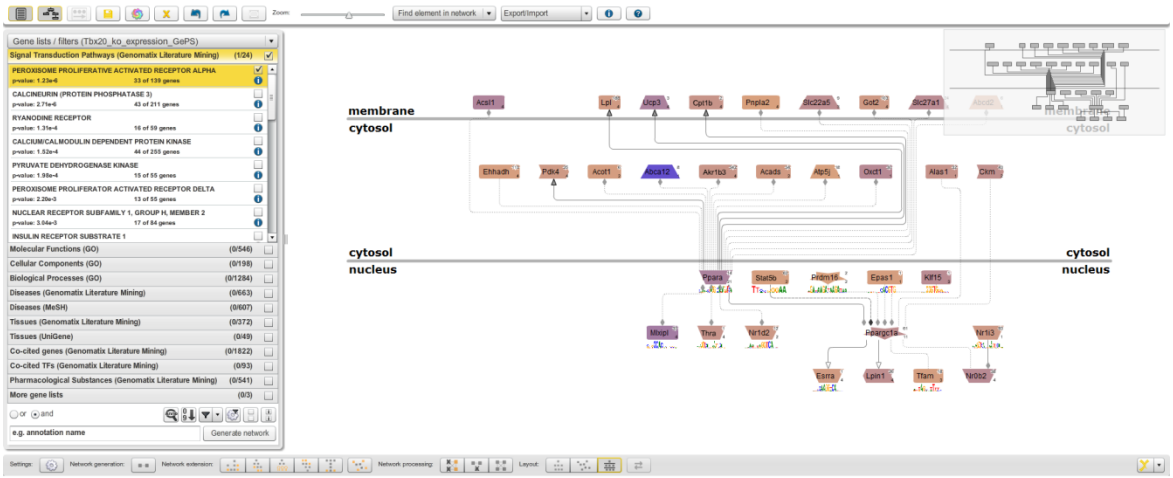
With the Genomatix Pathway System (GePS), you can generate gene networks and identify the biology that is overrepresented in a set of genes. Depending on the organism, there is a selection of biological categories, e.g. signal transduction pathway associations, GeneOntology (GO), diseases, and tissues.

From the *Differential Expression Analysis* section, run the Genomatix Pathway System for the down-regulated genes. To do this, remove the number from the field for the up-regulated genes, and change the entry for the down-regulated genes to 2300 to include all of them; then press the *Go* button.

In the output, you'll find lists of overrepresented terms in the different categories based on the Gene ID list you uploaded.

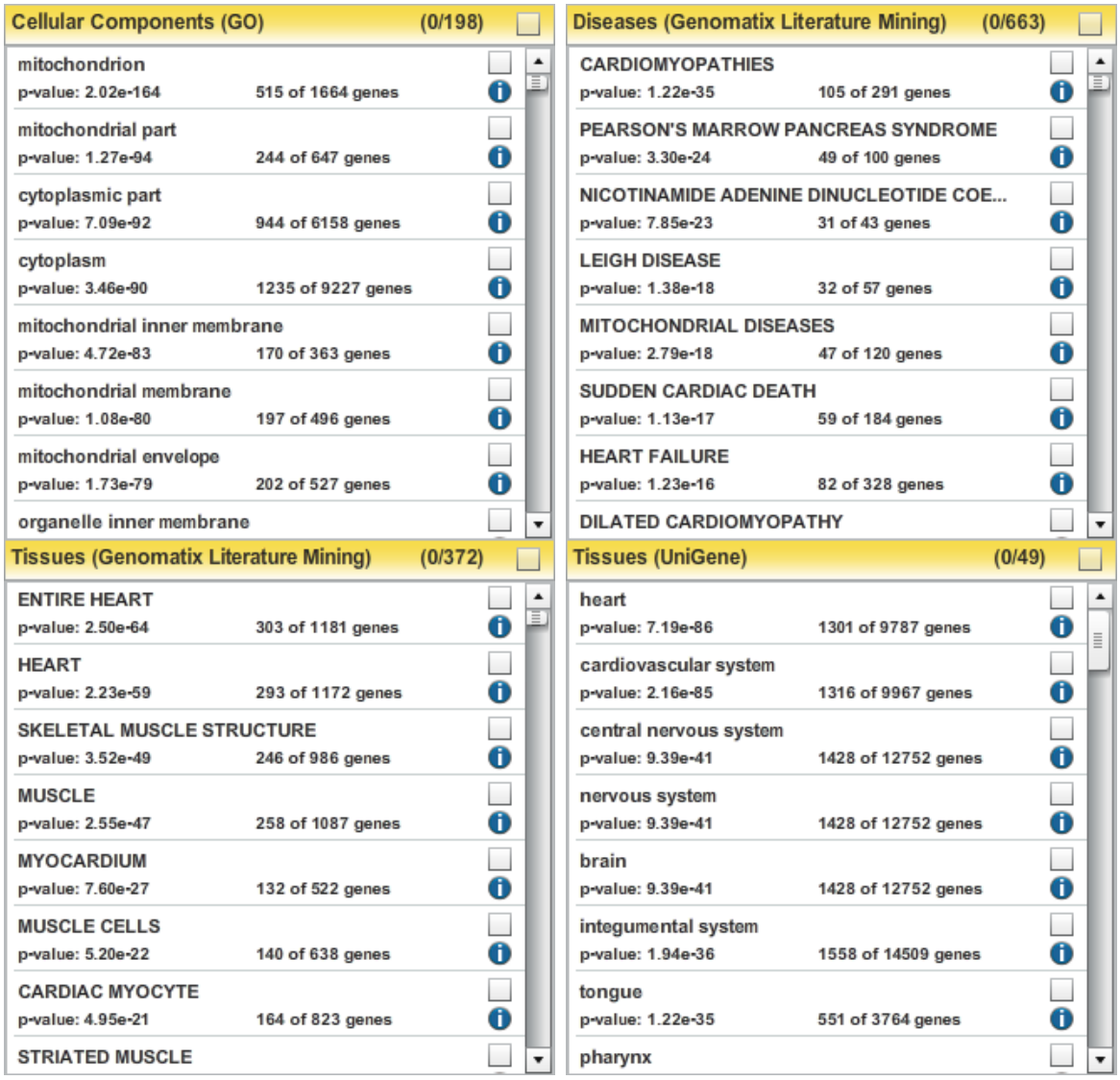
The top enriched literature mining based pathway is PPAR alpha, which plays an important role in heart physiology.

Click on the first entry to display the corresponding literature-based gene network.



The input genes are shown with a orange (weak down-regulation) to blue (strong down-regulation) colored background. Details will be shown during the workshop.

Other overrepresented biological annotations include *mitochondrion* in the *GO Cellular Components* category, *cardiomyopathies* among the *literature-mining based diseases*, and *heart tissue* based both on *literature mining* and *UniGene tissue* annotation.



Chip-sequencing analysis

ChIP-Seq workflow: regions bound by Tbx20 in the adult mouse heart

In the next example, you will learn how to analyze ChIP-Seq data, including peak finding and TFBS analysis.

Available peak finding algorithms

As ChIP-Seq data are inherently noisy, clustering of mapped ChIP-Seq reads is a prerequisite step for their analysis. Clustering algorithms use a distribution model of the reads for separating signal from noise.
Three different algorithms are available in RegionMiner for cluster detection in ChIP-Seq data: NGS Analyzer, and the public algorithms MACS (Model based Analysis for ChIP-Seq) and SICER (Spatial clustering for Identification of ChIP-Enriched Regions).

NGS Analyzer was developed by Genomatix; it identifies local enrichments (clusters) representing genomic regions bound by protein (ChIP-Seq) or being expressed (RNA-Seq). By default, the threshold applied by the clustering algorithm takes the density of the data set into account, assuming a Poisson distribution.
A control data file can be provided. A quantitative comparison of the clustered reads in the experimental data file to the reads in corresponding regions in the control file uses the Audic-Claverie algorithm (Audic & Claverie, 1997).

MACS was originally designed specifically for clustering of ChIP-Seq data with narrow peaks as you typically get from transcription factor binding. It uses a sliding window approach and assumes a Poisson distribution of the reads just as NGS Analyzer does. However, it uses a peak model generated from high confidence read cluster regions in the data to shift the reads to the assumed center of a protein binding region. It also uses the local read density background for peak calling, which NGS Analyzer does not do. MACS comes with its own quantitative background subtraction method against a control file.
MACS has been developed at the Dana-Farber Cancer Institute (Zhang et al, 2008). The GGA provides both versions 1.4 and 2 of the MACS implementation; the latter can also be used for broader peaks.

SICER (Zang et al., 2009) is particularly recommended for the analysis of histone modifications, which form broad peaks. It scores non-overlapping windows (typically of nucleosome length) based on the read count, assuming a Poisson distribution. Windows are flagged eligible based on a read count significance threshold, and adjacent eligible windows are grouped as islands (peaks). Small gaps of ineligible windows can be allowed within islands. The island score is the sum of the scores of the eligible windows in the island.

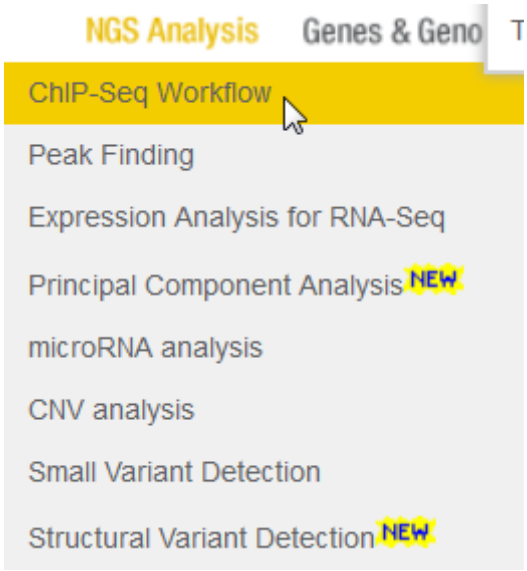
In the first step of the analysis, we will identify genomic regions bound by Tbx20 in wild type adult mouse heart, and run some downstream analyses on these ChIP peak regions.

For this we will use the Chip-Seq workflow, which is an automated process that includes a number of analyses: peak finding, read and peak classification, creation of a peak sequence file, and TFBS overrepresentation analysis.

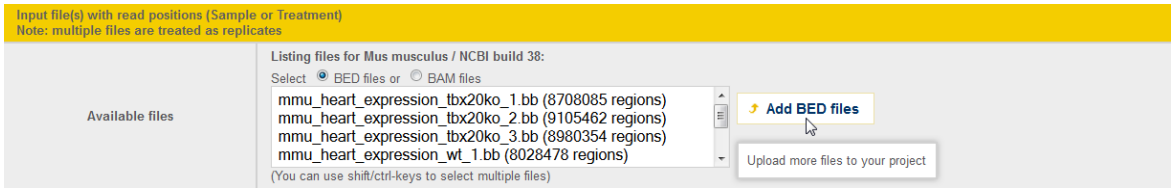
Additionally, a de novo definition of TF binding sites from the ChIP cluster sequences is possible. This uses the program CoreSearch, which can, of course, also be run separately.

The raw sequence tags from the experiment have been mapped to the human genome using the GMS. You find the files once more in the folder *workbench_home/Demo/NGS_Seminar/mmu_heart* on the GGA.

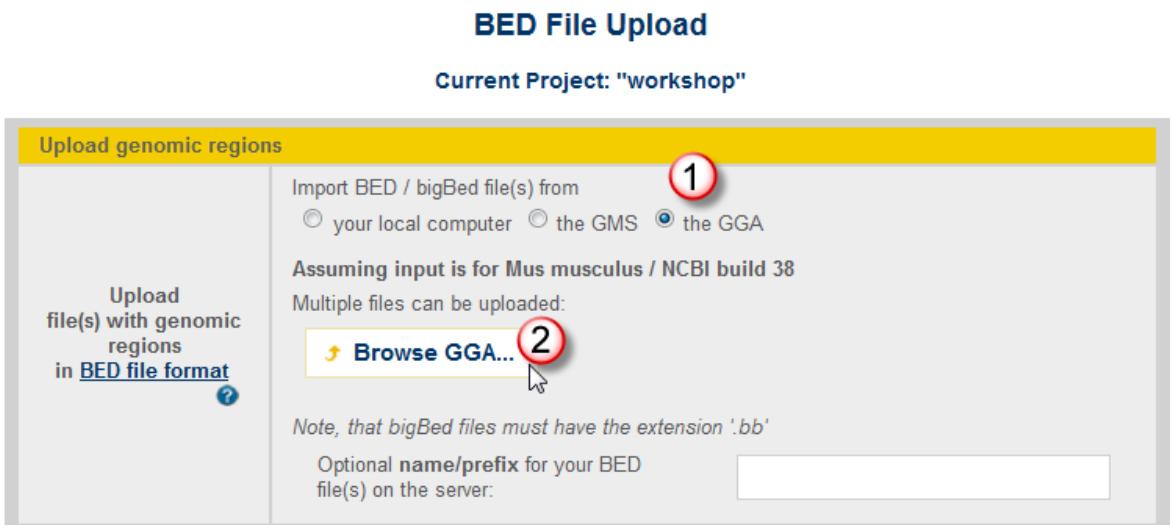
Please open the Genomatix Genome Analyzer in your browser, and select “ChipSeq Workflow” in the NGS Analysis menu.



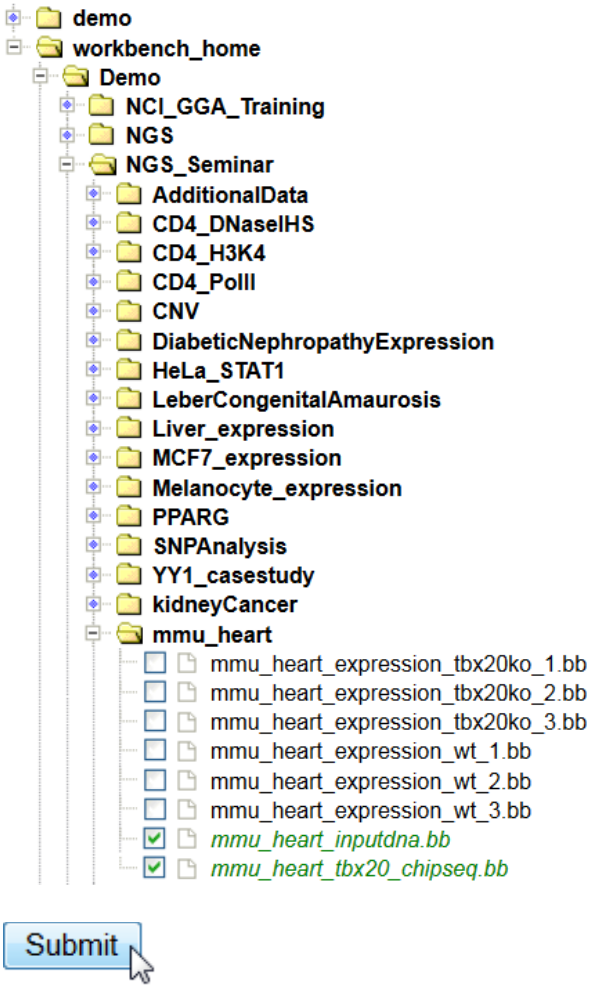
On the input page, press the Add BED files button.



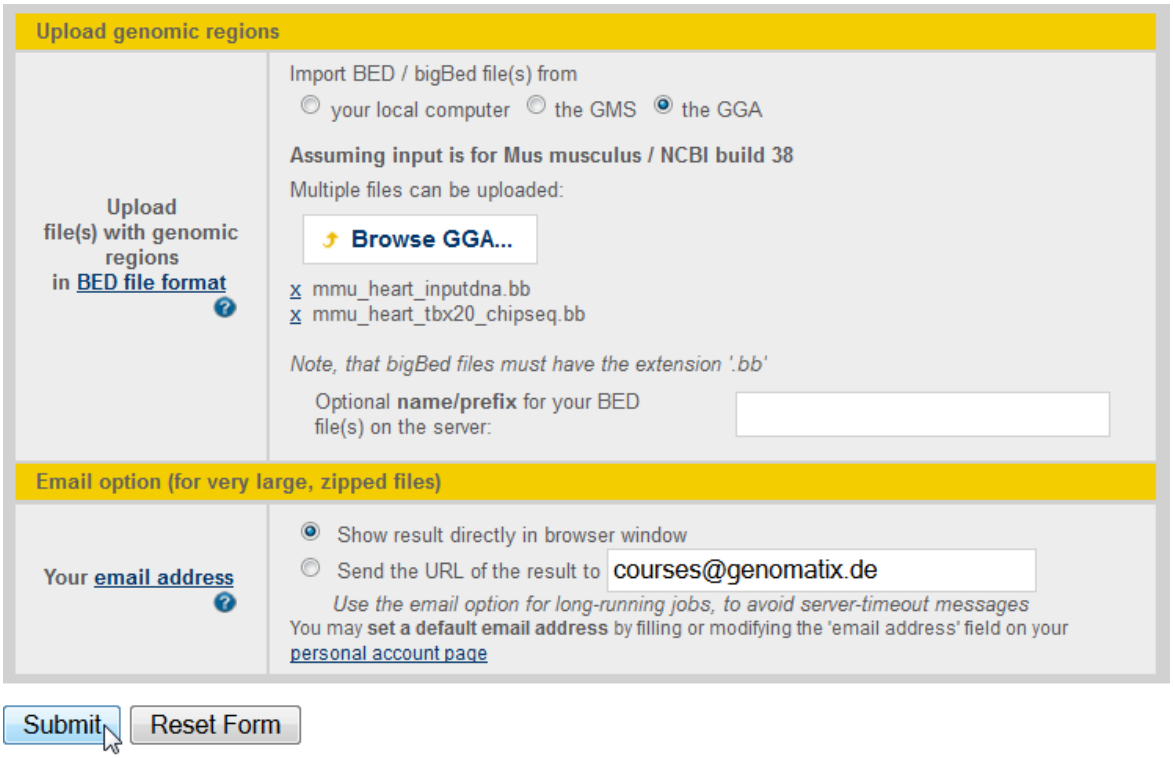
In the upload dialog, select the GGA for the file import and press the *Browse GGA* button.



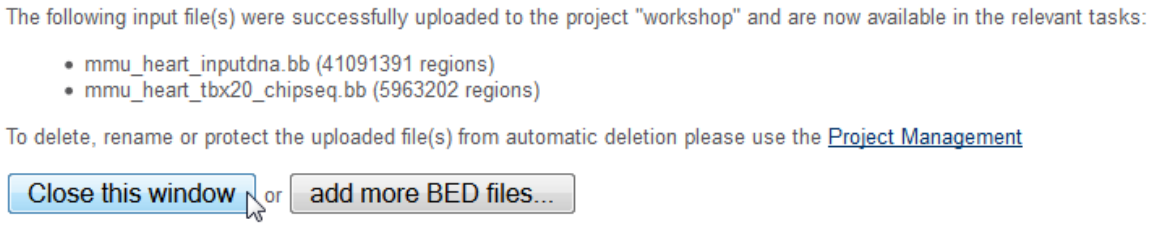
Select the last two files (input DNA and Tbx20 ChIP-Seq) in the folder *workbench_home/Demo/NGS_Seminar/mmu_heart*, and click on *Submit*.



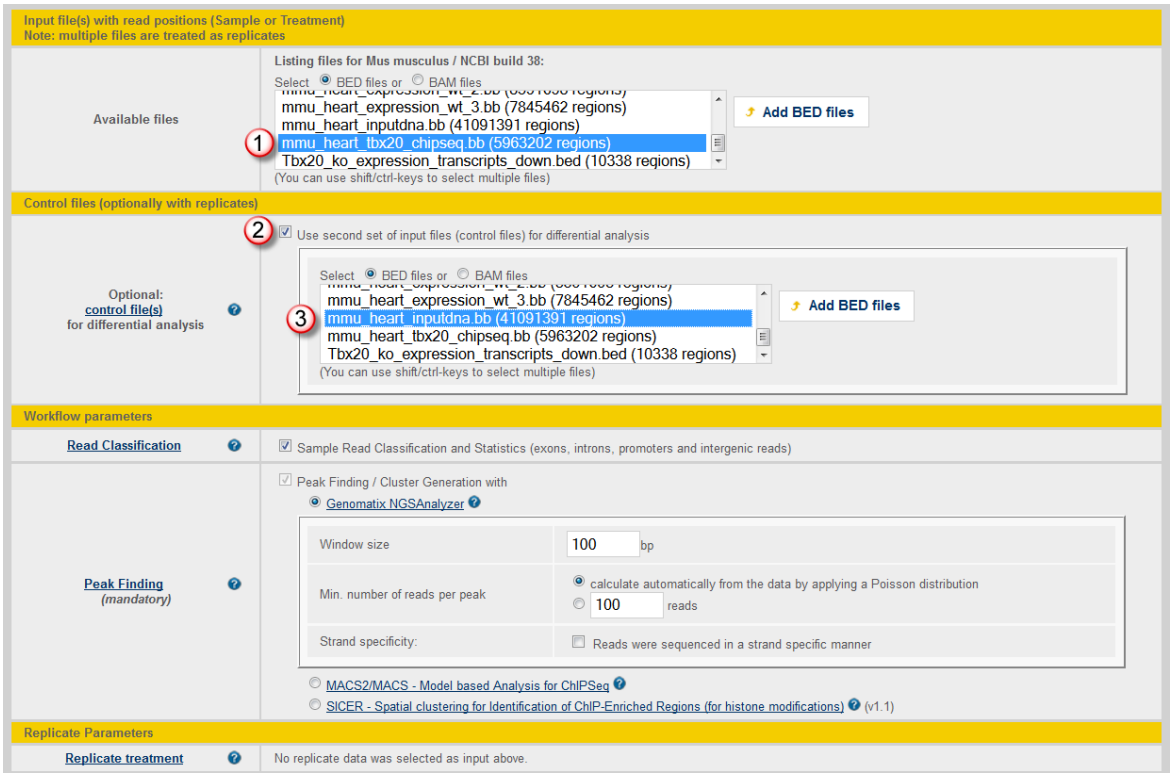
Press Submit in the upload dialog to start the import process.



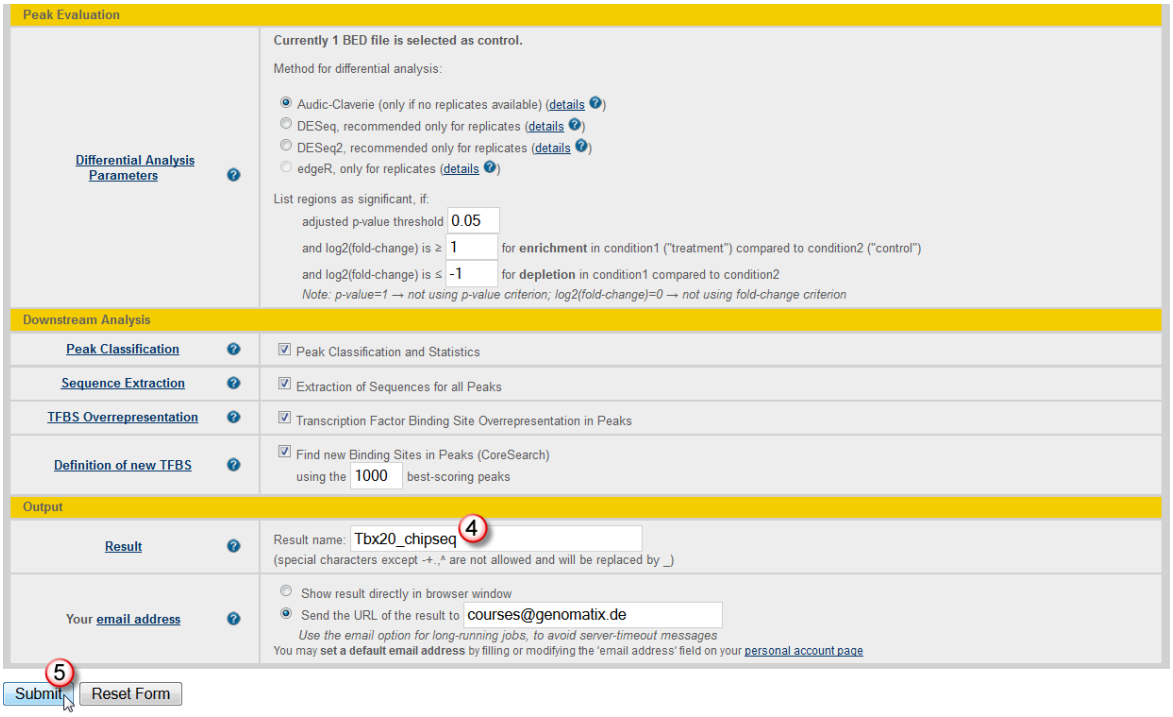
When the upload has finished, press the Close this window button.



In the *Available files* list, choose the Tbx20 ChIPSeq data set as treatment file. Activate the option *Use second set of input files...* and select the input DNA data set as control file. Please leave the workflow parameters at the default values.

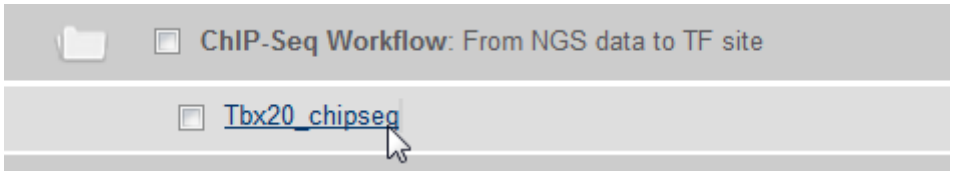


In this example, we'll also use the defaults of the peak evaluation and downstream analysis parameters. Please provide a result name, and start the analysis with the standard e-mail option.



You'll see a message informing you that the job has been started.

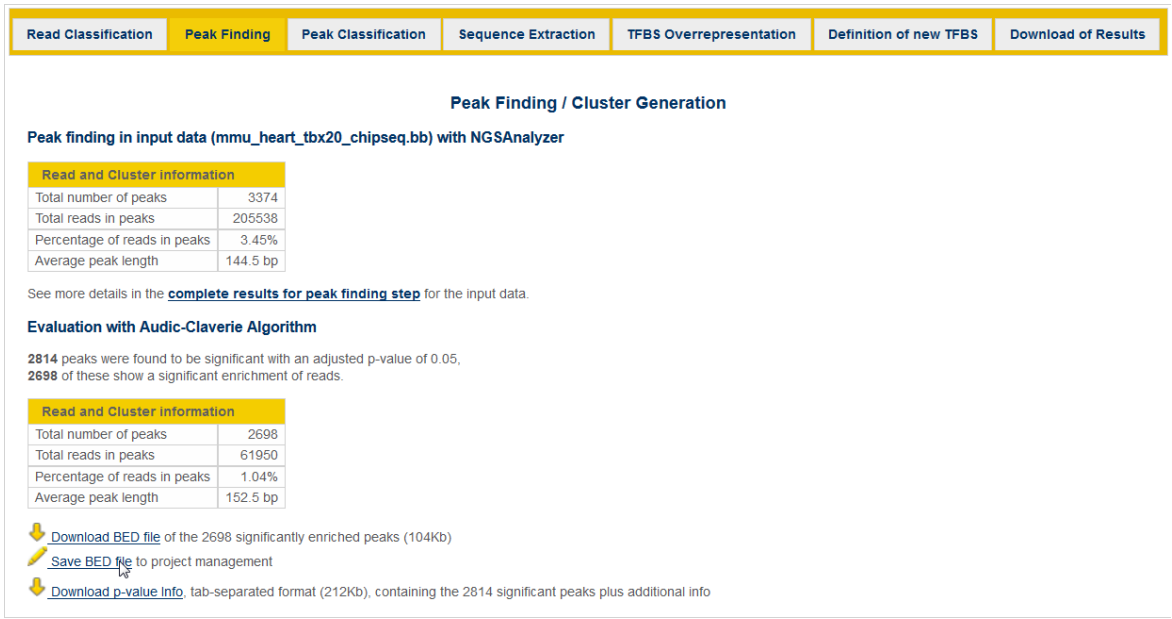
When the job has finished, open your project folder and the result group "ChIP-Seq Workflow" and click on the entry to open the result.



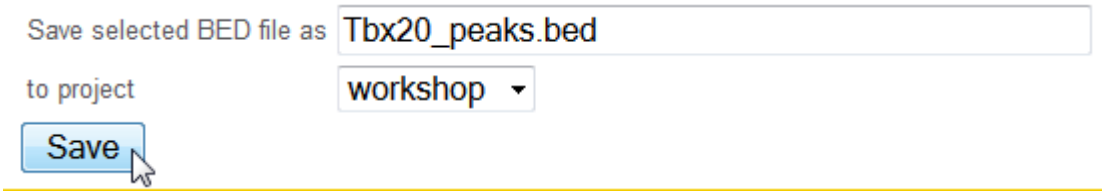
Peak finding

The output page has its own navigation bar, which is used to access each workflow result. The peak finding result is shown by default.

In the experimental sample, 3374 peaks were found originally, of which 2698 enriched peaks remain after Audic-Claverie evaluation. 1.04% of the reads are in peaks, which is relatively low.



Please save the BED file with significantly enriched clusters to the project management.

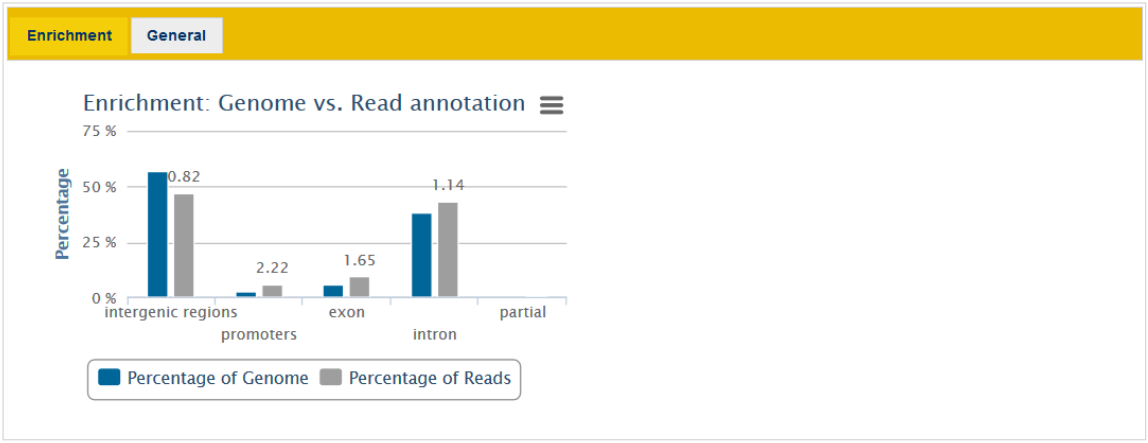


Read classification

The read classification in shows that enrichment in promoters is only slightly higher for the Tbx20 ChIP-Seq reads than for the input control:

Read Classification on mmu_heart_tbx20_chipseq.bb

General Statistics	
Total number of Reads:	5963202
Total basepairs:	214675272
Minimum Read length:	36
Maximum Read length:	36
Average Read length:	36.0



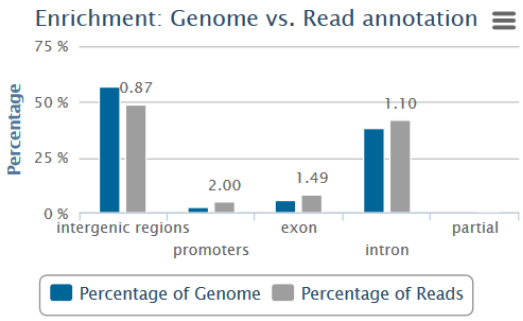
Type of genomic element	Number of Reads	Percentage of Reads	Percentage in Genome	Enrichment compared to Genome
Exon	560195	9.4%	5.7%	1.6
Partial	62348	1.0%	-	-
Intron	2569635	43.1%	37.8%	1.1
Intergenic regions	2771024	46.5%	56.5%	0.8
Sum of above	**5963202**	**100.0%**	**-**	**-**
Promoters	360351	6.0%	2.7%	2.2

Distribution of Reads on the Genome
>>> show details <<<

Read Classification on mmu_heart_inputdna.bb

General Statistics	
Total number of Reads:	41091391
Total basepairs:	1479290076
Minimum Read length:	36
Maximum Read length:	36
Average Read length:	36.0

Enrichment | General



Type of genomic element	Number of Reads	Percentage of Reads	Percentage in Genome	Enrichment compared to Genome
Exon	3489333	8.5%	5.7%	1.5
Partial	379444	0.9%	-	-
Intron	17134331	41.7%	37.8%	1.1
Intergenic regions	20088283	48.9%	56.5%	0.9
Sum of above	**41091391**	**100.0%**	**-**	**-**
Promoters	2206277	5.4%	2.7%	2.0

Distribution of Reads on the Genome

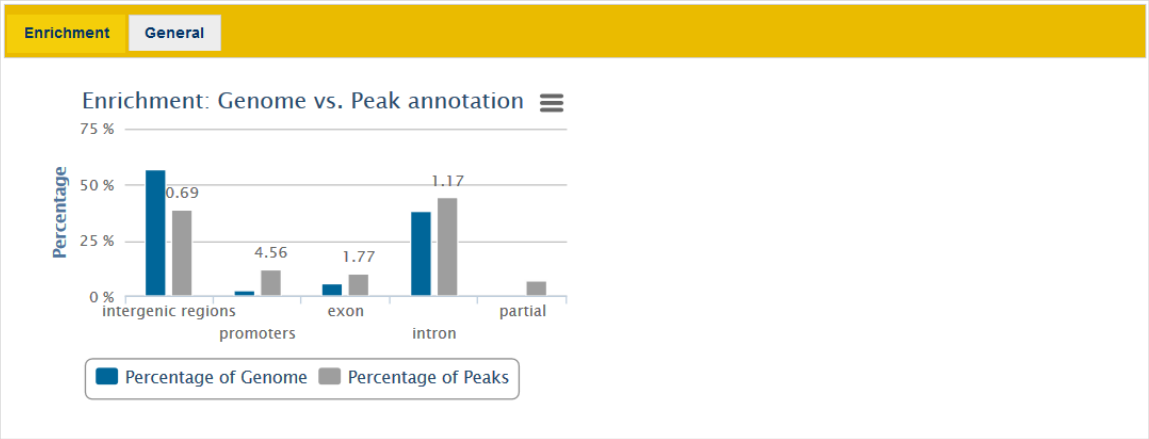
>>> show details <<<

Peak classification

The enrichment in promoters is 4.56 fold for peaks, approximately double of that for reads.

Peak Classification on claverie_result.bed

General Statistics	
Total number of Peaks:	2698
Total basepairs:	411423
Minimum Peak length:	36
Maximum Peak length:	4808
Average Peak length:	152.5



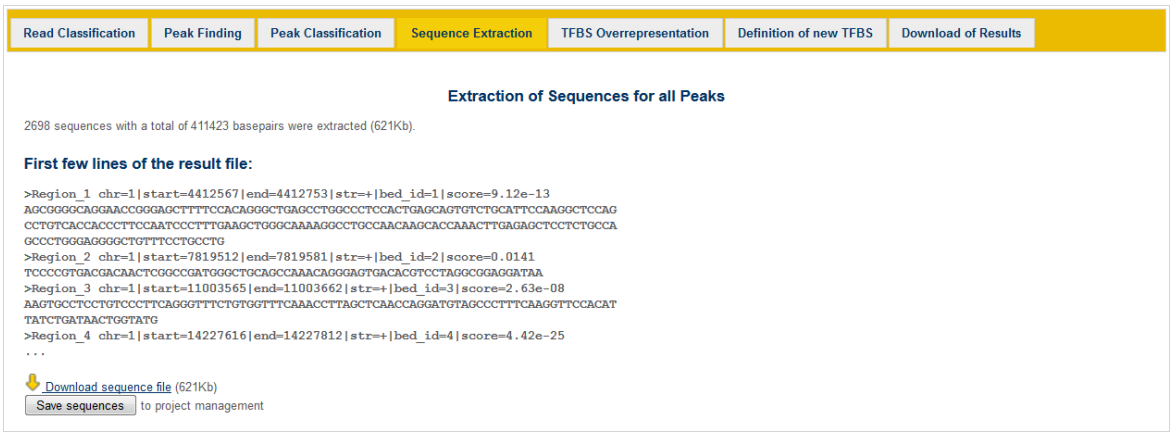
Type of genomic element	Number of Peaks	Percentage of Peaks	Percentage in Genome	Enrichment compared to Genome
Exon	272	10.1%	5.7%	1.8
Partial	183	6.8%	-	-
Intron	1190	44.1%	37.8%	1.2
Intergenic regions	1053	39.0%	56.5%	0.7
Sum of above	**2698**	**100.0%**	**-**	**-**
Promoters	332	12.3%	2.7%	4.6

Distribution of Peaks on the Genome

>>> show details <<<

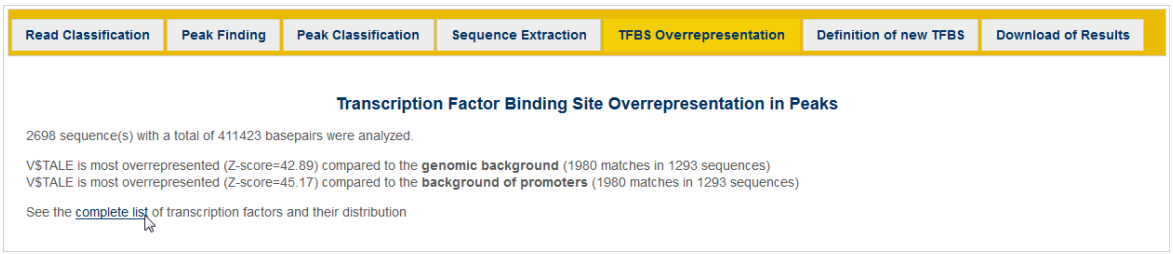
Sequence extraction

The peak sequences can be saved in the next section:



TFBS overrepresentation

Next, we'll have a look which transcription factor binding sites can be found in the clusters. A short summary of the TFBS analysis is given in the overview: V$TALE is most overrepresented, both against a genomic and a promoter background.



Click the “complete list” link to open the detailed result page.

You'll see some statistics on top and then a table containing all transcription factor binding site matches together with overrepresentation values and Z-scores. V$BRAC, the binding site family for Tbx20, ranks second after V$TALE in the overrepresentation.

Listing of all TF Families

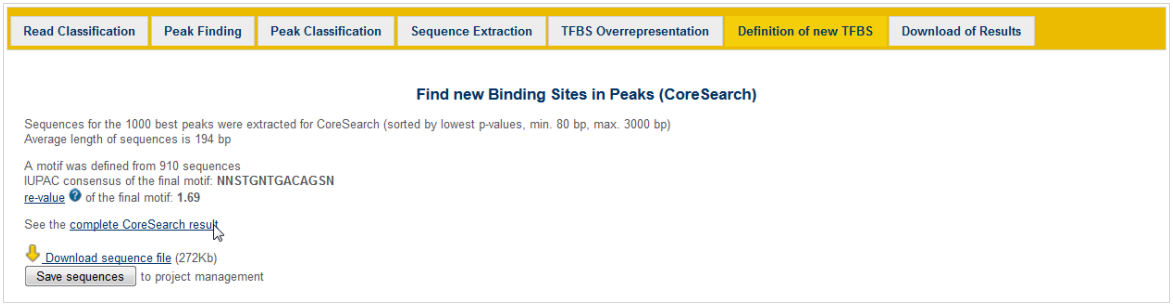
TF Families	Prom. assoc. known	Nr. of Input Seq. with Match	Nr. of Matches in Input	Match details	Expected (genome) ± Std.dev.	Over representation (genome)	Z-Score (genome)	Expected (promoters) ± Std.dev.	Over representation (promoters)	Z-Score (promoters)
V$TALE	no	1293	1980	list/seq	781.52±27.93	2.53	42.89	746.54±27.30	2.65	45.17
V$BRAC	no	1444	2198	list/seq	944.36±30.70	2.33	40.83	873.14±29.52	2.52	44.87
V$MYOD	no	913	1782	list/seq	834.02±28.85	2.14	32.84	1028.98±32.04	1.73	23.49
V$NF1F	no	671	1000	list/seq	414.95±20.36	2.41	28.71	463.37±21.51	2.16	24.92
V$ZF5F	yes	137	318	list/seq	84.52±9.19	3.76	25.34	697.82±26.39	0.46	-14.41
V$ZF11	no	398	475	list/seq	157.44±12.55	3.02	25.27	240.11±15.49	1.98	15.13
V$CTCF	yes	524	725	list/seq	294.94±17.17	2.46	25.02	836.10±28.89	0.87	-3.86
V$AP4R	no	352	451	list/seq	151.13±12.29	2.98	24.36	228.73±15.12	1.97	14.67
V$AP1R	no	1197	2099	list/seq	1263.12±35.49	1.66	23.54	1271.20±35.60	1.65	23.24
V$AP2F	yes	439	708	list/seq	309.95±17.60	2.28	22.59	676.47±25.99	1.05	1.19
V$NRF1	yes	157	313	list/seq	99.57±9.98	3.14	21.34	589.03±24.25	0.53	-11.40
V$E2FF	yes	725	1239	list/seq	695.03±26.34	1.78	20.63	1732.38±41.53	0.72	-11.89
V$MYRF	no	267	285	list/seq	89.46±9.46	3.19	20.62	117.76±10.85	2.42	15.37
O$MTEN	yes	174	226	list/seq	67.62±8.22	3.34	19.20	328.87±18.13	0.69	-5.70
V$HDBP	yes	60	91	list/seq	15.57±3.95	5.84	18.99	186.21±13.64	0.49	-7.02

The list is sorted by the Z-score of the overrepresentation over the genome. The overrepresentation for V$BRAC is about 2.3 - 2.5 fold over genome and promoter background, respectively, and the Z-scores are quite high, indicating that it is statistically highly unlikely to find such an overrepresentation. You can click any column header to sort by that column; repeated clicking inverts the sort order.

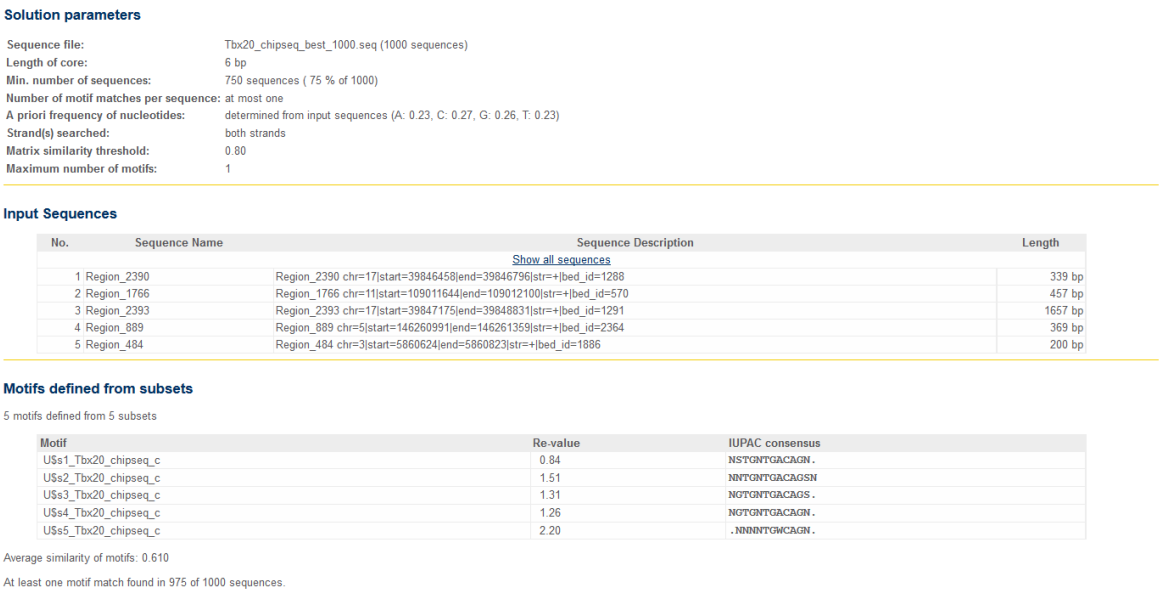
Definition of new TFBS

The TFBS overrepresentation analysis uses pre-defined binding site matrices from the MatBase/MatInspector library provided with the Genomatix Genome Analyzer. It is, however, also possible to define your own matrices from the data generated by the ChIP-Seq experiment. In the workflow, the Tbx20 cluster sequences were submitted to CoreSearch to generate a new Tbx20 binding site matrix.

The next item in the workflow output overview is the CoreSearch result. The sequences of all clusters were used to generate a new matrix. The IUPAC consensus of the defined motif is shown. For details, please click the “complete CoreSearch result” link.



Here is an outline of the CoreSearch algorithm: as a first step, CoreSearch randomly picks sets of 100 input sequences to generate 5 matrices, which are grouped into a family. The IUPAC sequences of the matrices are displayed in the output below the list of input sequences:



Solution parameters

Sequence file:	Tbx20_chipseq_best_1000.seq (1000 sequences)
Length of core:	6 bp
Min. number of sequences:	750 sequences (75 % of 1000)
Number of motif matches per sequence:	at most one
A priori frequency of nucleotides:	determined from input sequences (A: 0.23, C: 0.27, G: 0.26, T: 0.23)
Strand(s) searched:	both strands
Matrix similarity threshold:	0.80
Maximum number of motifs:	1

Input Sequences

No.	Sequence Name	Sequence Description	Length
		Show all sequences	
1	Region_2390	Region_2390 chr=17|start=39846458|end=39846796|str=+|bed_id=1288	339 bp
2	Region_1766	Region_1766 chr=11|start=109011644|end=109012100|str=+|bed_id=570	457 bp
3	Region_2393	Region_2393 chr=17|start=39847175|end=39848831|str=+|bed_id=1291	1657 bp
4	Region_889	Region_889 chr=5|start=146260991|end=146261359|str=+|bed_id=2364	369 bp
5	Region_484	Region_484 chr=3|start=5860624|end=5860823|str=+|bed_id=1886	200 bp

Motifs defined from subsets

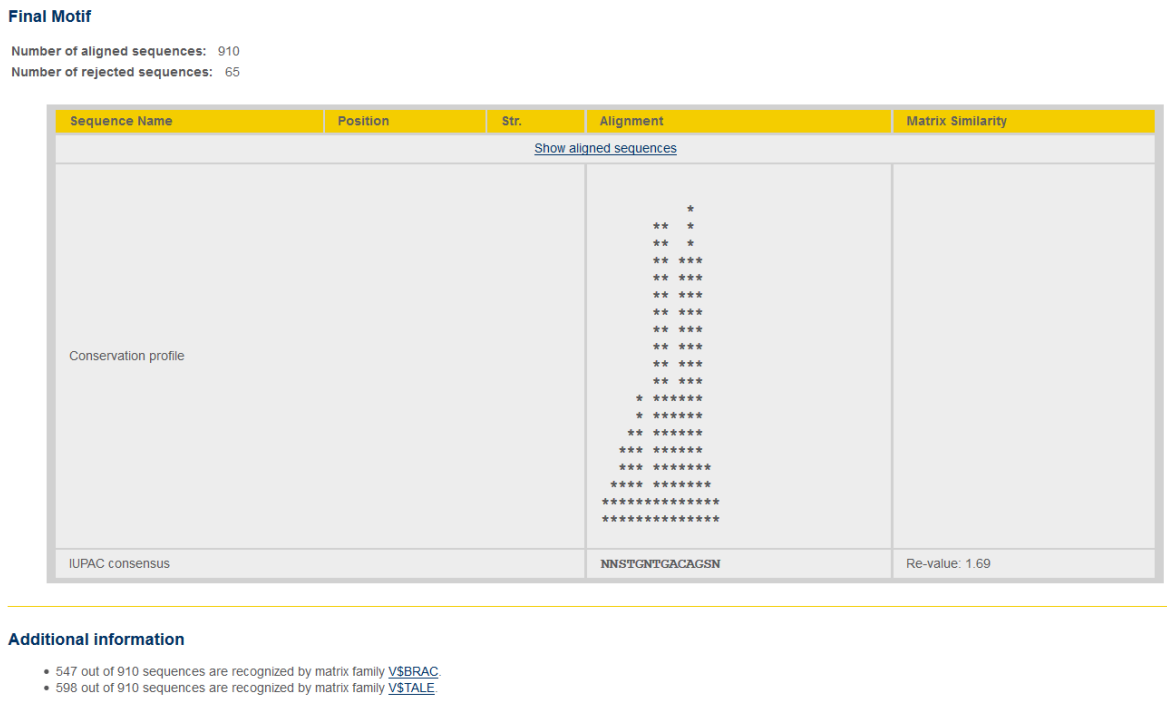
5 motifs defined from 5 subsets

Motif	Re-value	IUPAC consensus
U$s1_Tbx20_chipseq_c	0.84	NSTGNTGACAGN.
U$s2_Tbx20_chipseq_c	1.51	NNTGNTGACAGSN
U$s3_Tbx20_chipseq_c	1.31	NGTGNTGACAGS.
U$s4_Tbx20_chipseq_c	1.26	NGTGNTGACAGN.
U$s5_Tbx20_chipseq_c	2.20	.NNNNTGWCAGN.

Average similarity of motifs: 0.610

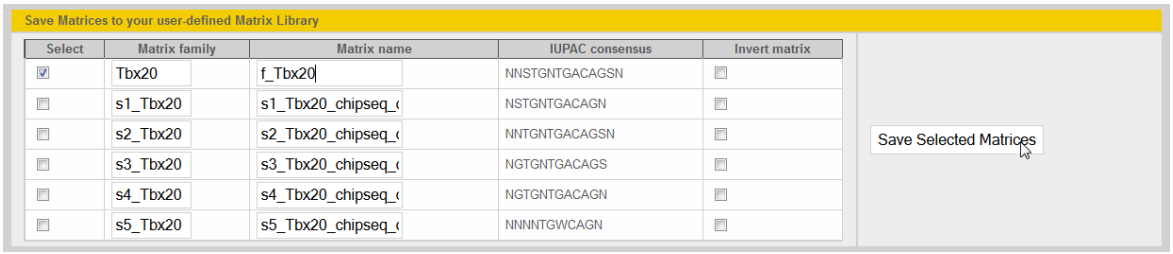
At least one motif match found in 975 of 1000 sequences.

All input sequences are then scanned for matches to the new matrix family, and the best match of each sequence is used to generate the final matrix. Its conservation profile is displayed at the end of the output page.

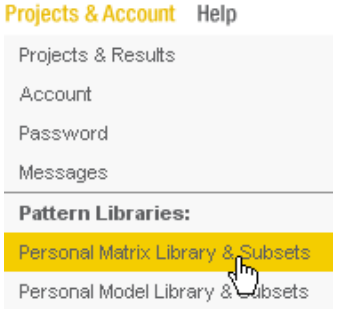


A bit more than half of the sequences used for generation of the matrix are also recognized by the existing V$BRAC matrix family.

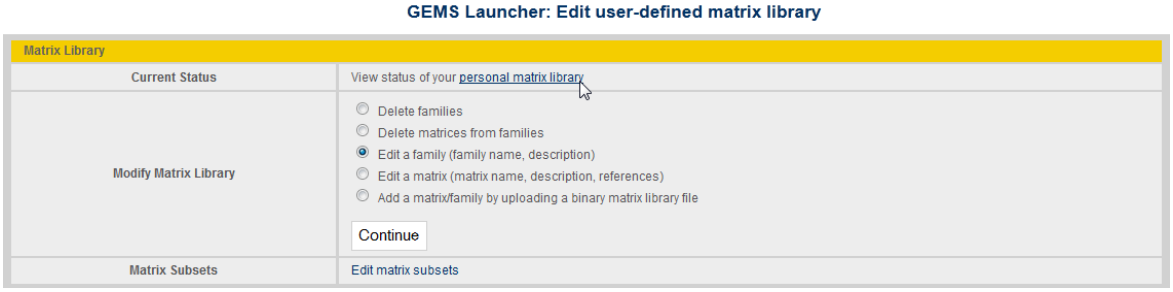
You can save any of the new matrices (the final one as well as the five matrices generated in the first step) in the 'Save Matrices to your user-defined Matrix Library' section at the bottom of the page. They are then available in tools applying matrix searches, such as MatInspector.



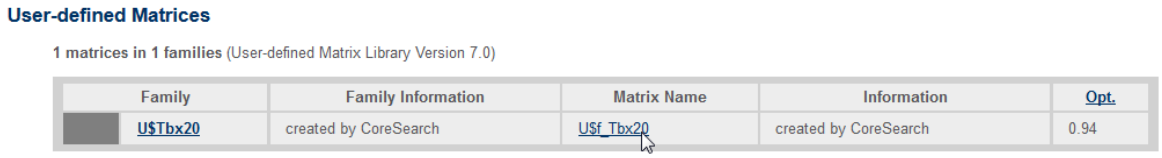
You can view your new matrices if you click the ‘Personal Matrix Library’ link in the menu:



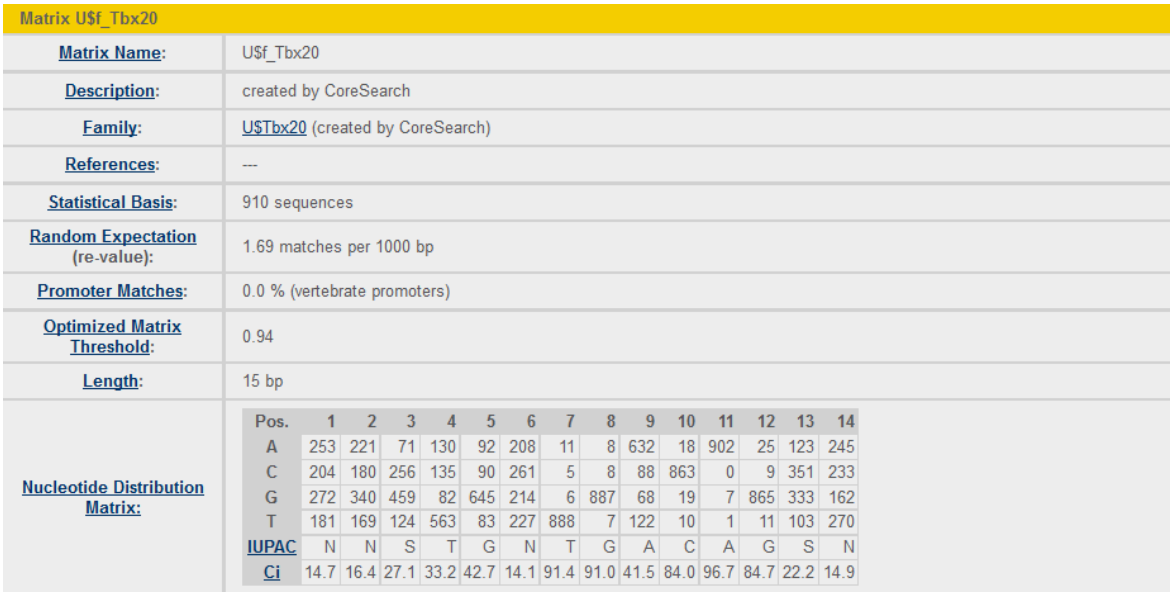
Select the “personal matrix library” link as shown below:



Click on the first matrix name to display detailed information for this matrix.

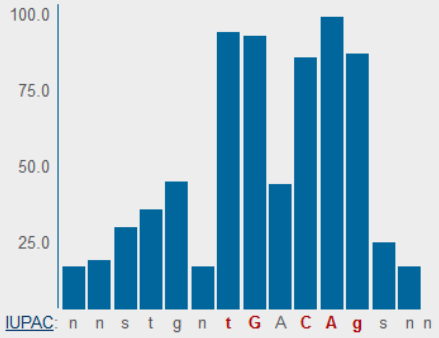


You'll see some statistics and the nucleotide distribution including IUPAC translation and consensus index for each position, which is a measure for conservation.



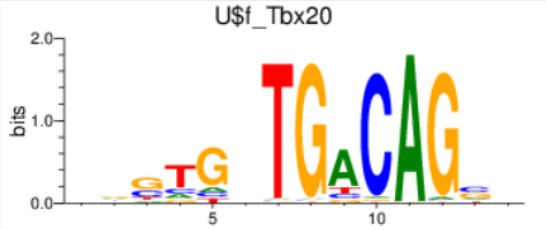
The conservation profile of the binding site definition is also shown in a column chart and as a sequence logo.

Profile:



- Basepairs marked red show a high information content, i.e. the matrix exhibits a high conservation (ci-value > 60) at this position.
- Basepairs in capital letters denote the core sequence used by MatInspector.

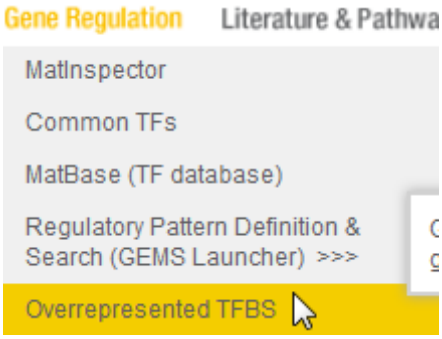
Sequence Logo:



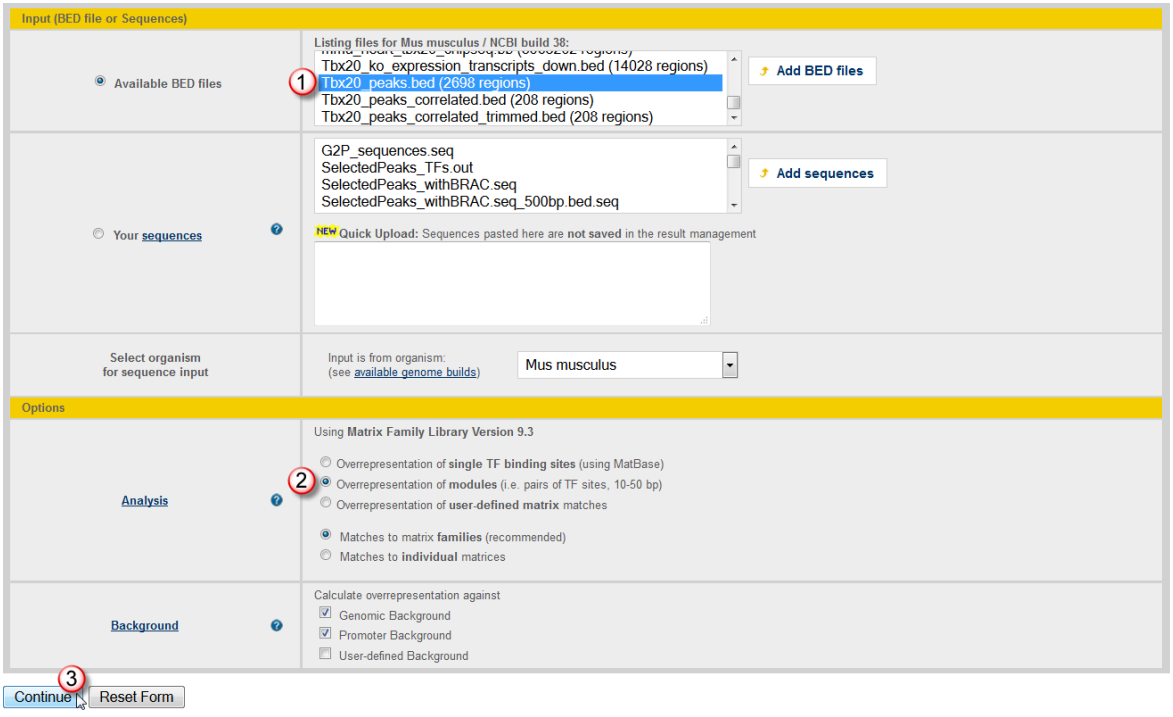
TFBS module overrepresentation

The TFBS overrepresentation analysis in the ChIP-Seq workflow considers only single binding site matches. As TFs often work in concert, it makes sense to analyze the ChIP regions for combinations of binding sites that could represent transcriptional modules, or parts thereof. Let's see if there are any combinations with other binding sites that can be found more often than others in our Tbx20 peaks.

Please select “Overrepresented TFBS” from the Gene Regulation menu

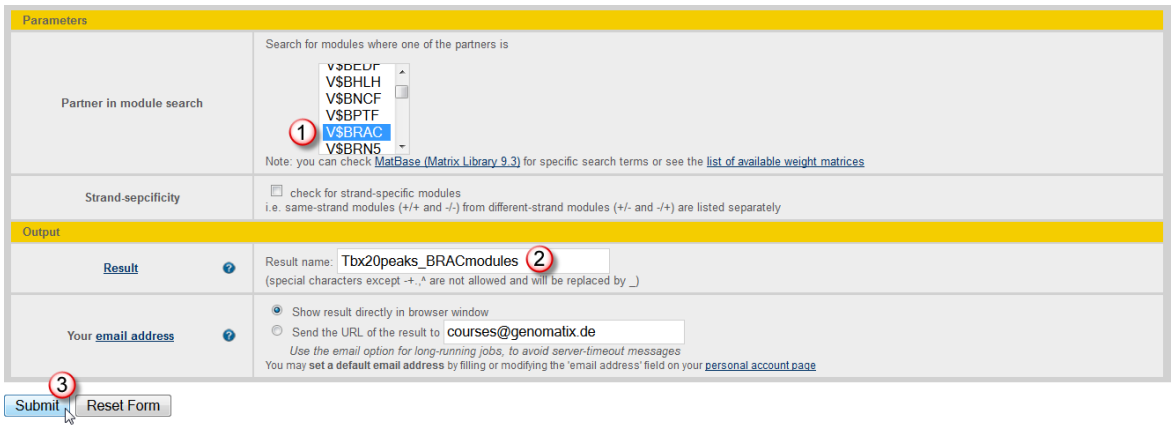


On the input page, select the Tbx20 peak file you saved on the ChIP-Seq workflow output in the list of previously uploaded BED files.



In the “options” section, click the radio button next to “Module overrepresentation (i.e. pairs of TF sites, 10-50 bp)”, and continue.

On the next page, choose one TF binding site family as a partner for searching for modules. Otherwise the number of possible combinations would be too high to calculate meaningful results in appropriate time. Of course, we choose the 'V$BRAC' family (containing transcription factor binding sites for Tbx20 matrices). Provide a result name, and press the Submit button.



This is the start of the output list:

Listing of all Modules with V$BRAC

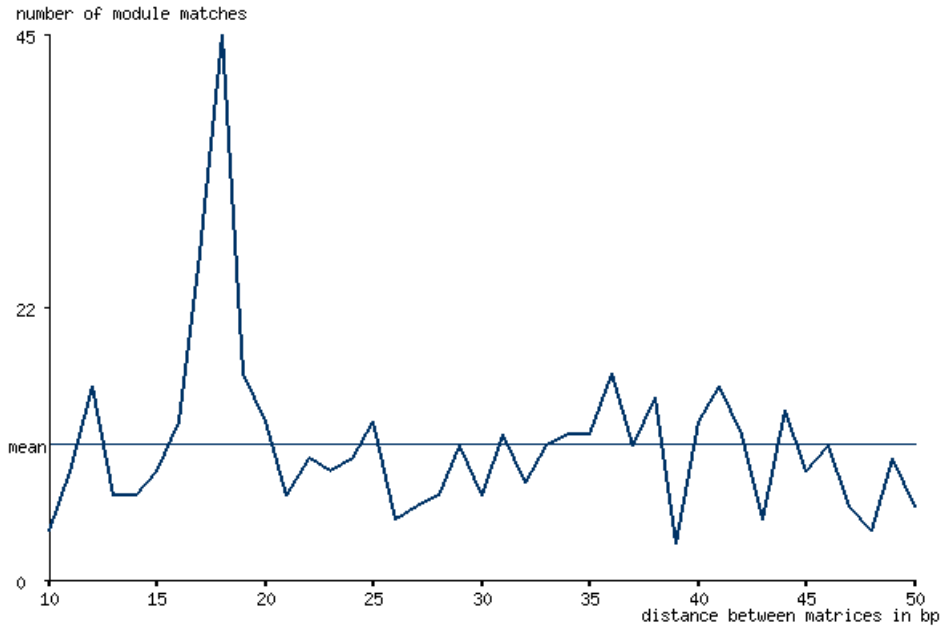
Modules with V$BRAC	Distance Score	Prom. assoc. known	Nr. of Input Seq. with Match	Nr. of Matches in Input	Match details	Expected (genome) ± Std.dev.	Over representation (genome)	Z-Score (genome)	Expected (promoters) ± Std.dev.	Over representation (promoters)	Z-Score (promoters)
V$BRAC-V$TALE	2.161	no	490	879	list	184.63±13.58	4.76	51.08	150.68±12.27	5.83	59.30
V$BRAC-V$MYOD	2.802	no	355	843	list	188.02±13.71	4.48	47.74	201.72±14.20	4.18	45.13
V$AP1R-V$BRAC	2.024	no	472	908	list	249.56±15.79	3.64	41.66	230.86±15.19	3.93	44.55
V$BRAC-V$NF1F	4.847	no	286	459	list	84.76±9.21	5.42	40.60	84.39±9.19	5.44	40.73
V$AP4R-V$BRAC	2.142	no	142	235	list	31.77±5.64	7.40	35.97	41.23±6.42	5.70	30.10
V$BRAC-V$ZF11	1.858	no	156	216	list	27.51±5.24	7.85	35.84	40.40±6.36	5.35	27.55
V$BRAC-V$CTCF	3.152	yes	205	306	list	55.85±7.47	5.48	33.41	112.92±10.63	2.71	18.13
V$AP2F-V$BRAC	2.738	no	172	319	list	62.51±7.91	5.10	32.38	105.76±10.28	3.02	20.69
V$BRAC-V$EGRF	1.595	no	202	442	list	115.03±10.72	3.84	30.44	217.41±14.74	2.03	15.20
V$BRAC-V$P53F	3.862	no	237	476	list	129.60±11.38	3.67	30.39	121.81±11.03	3.91	32.05
V$BRAC-V$SP1F	2.244	no	236	391	list	99.45±9.97	3.93	29.19	186.04±13.64	2.10	14.99
V$BRAC-V$HAND	1.980	no	412	679	list	233.31±15.27	2.91	29.15	226.40±15.04	3.00	30.06
V$BRAC-V$NEUR	3.513	no	283	433	list	119.61±10.93	3.62	28.61	116.49±10.79	3.72	29.28
V$BRAC-V$EBOX	2.193	no	255	445	list	126.43±11.24	3.52	28.29	145.64±12.07	3.06	24.77

V$TALE, V$MYOD, V$AP1R, and V$NF1F are the most overrepresented partners of Tbx20 sites in modules consisting of two sites with a distance of 10 to 50 bp in between.

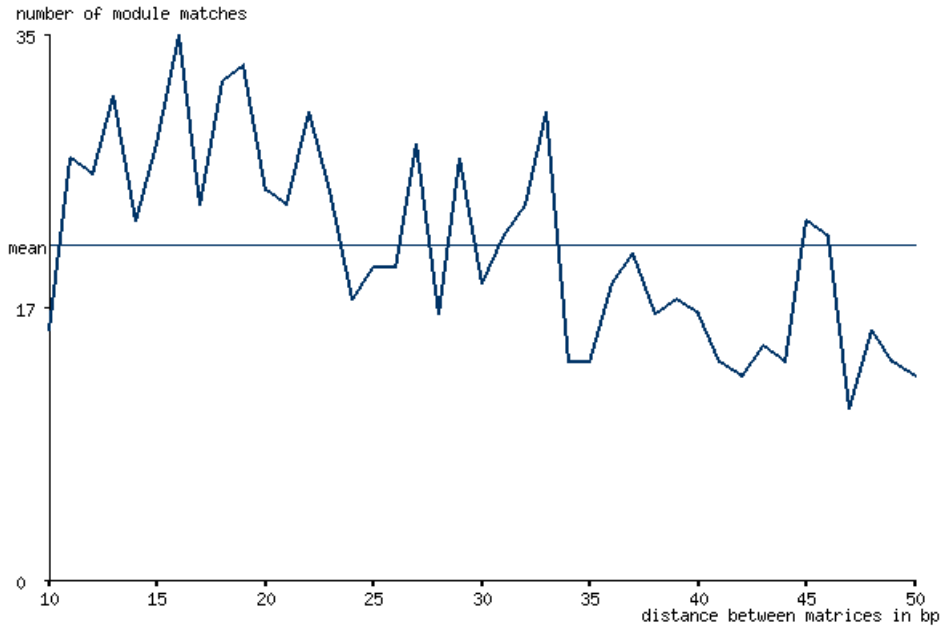
The distance score can be used for sorting module matches with one or a few preferred distances between the sites in the input sequences. A high score would indicate a strong distance preference.

To see a profile of the distribution of distances between the binding sites in any model, please click the corresponding *list* link in the *match detail* column.

The distance profile of the pair of BRAC-NF1F combinations, with a distance score of 4.847, clearly shows a peak at 183 bp over a low background.



In contrast, the top overrepresented combination of BRAC with TALE has lower distance score (2.161), and doesn't show a clear peak:



In summary, regions of Tbx20 binding sometimes show specific distance-conserved patterns of BRAC sites with other TF binding sites. The fraction of matches with preferred distances can be up to 20% of the total matches in the regions.

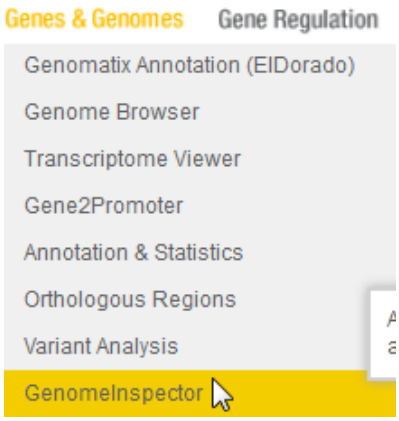
Integration of expression and ChIP-Seq data

Positional correlation of Tbx20 peaks with differentially expressed transcripts

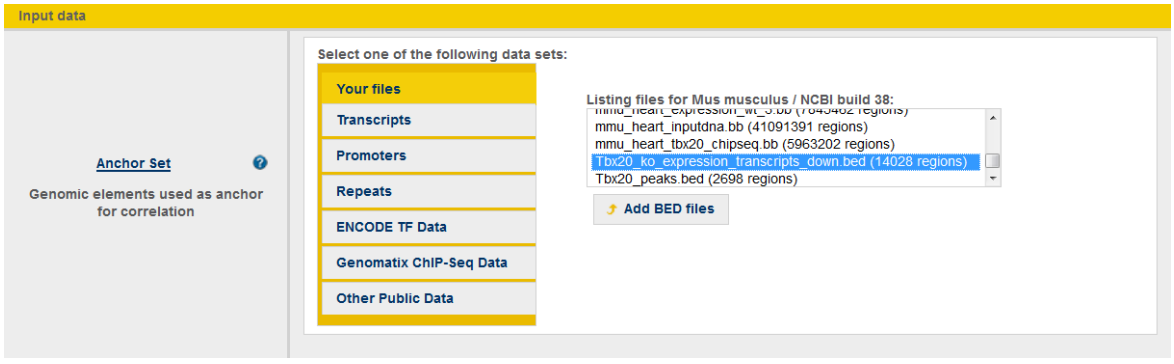
In the next step, we will predict which genes that are differentially expressed in Tbx20 knock-out mouse hearts are direct targets of Tbx20. For this, we will use the program GenomeInspector and visualize the positional correlation of the starts of the down-regulated transcripts with the Tbx20 ChIP peaks.

GenomeInspector uses one BED file as anchor set and, based on the genomic positions of the regions in the file, draws a correlation graph for up to 6 other BED files (the partner sets). The graph shows the summarized coverage with regions from the partner sets in the vicinity of the regions in the anchor set.

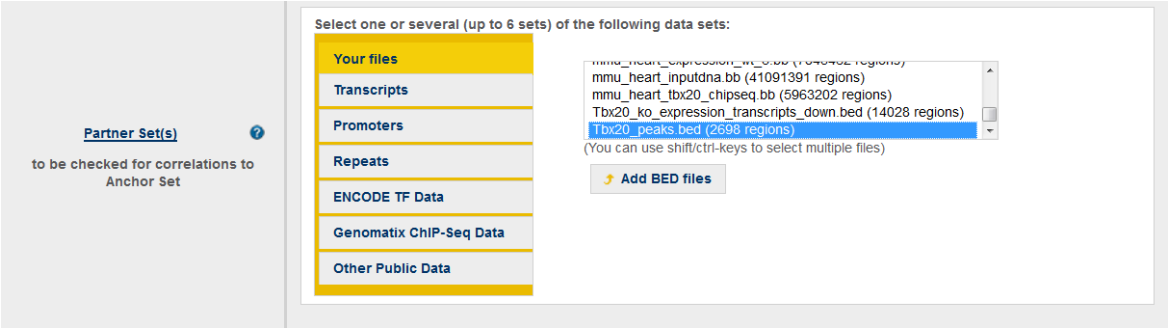
Please start GenomeInspector from the Gene & Genomes menu.



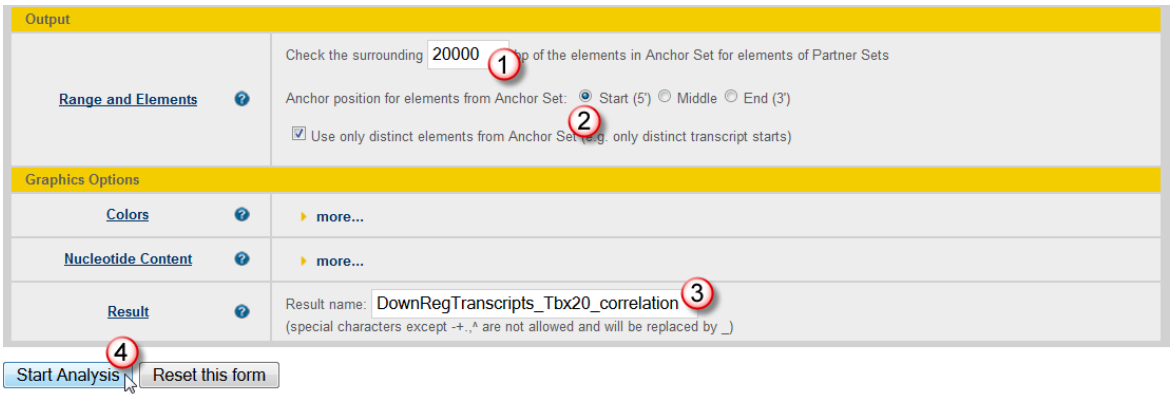
Select the BED file of down-regulated transcripts from your files in the anchor set list.



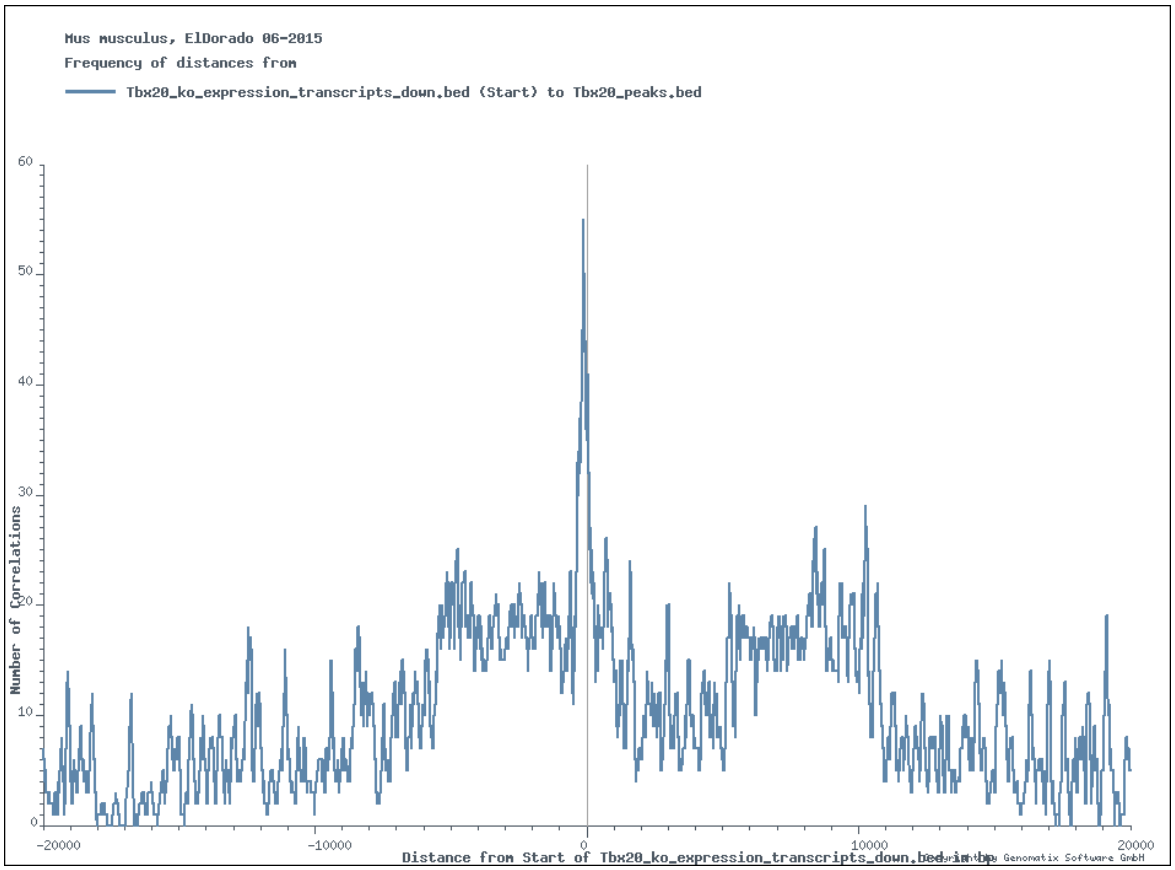
Select the Tbx20 peaks as partner set.



Set the range to the surrounding 20000 bps; in this way, also more distal regulatory regions will be included. Make sure the anchor position is at the start of the anchor set (i.e. the transcript starts), provide a result name, and start the analysis.



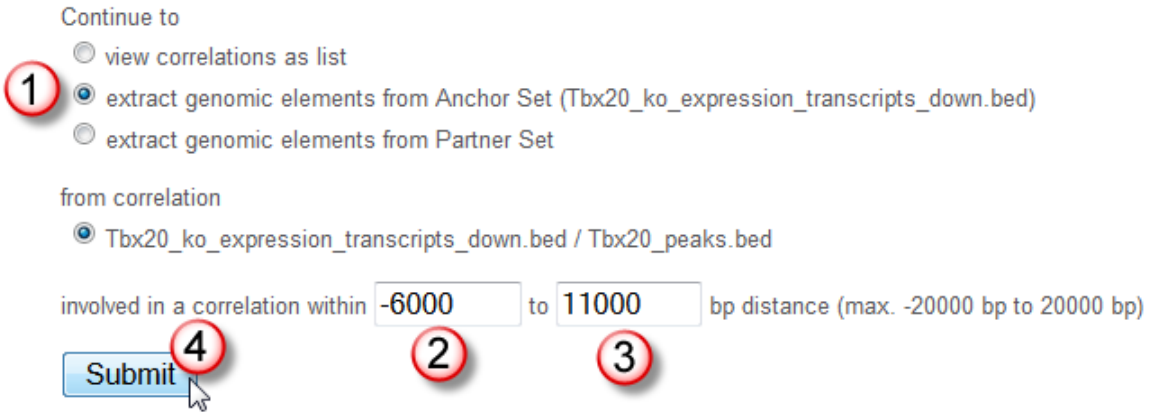
The graph shows a narrow peak around the transcript start sites, representing the region with the highest density of Tbx20 peaks. There is also a slightly elevated plateau ranging from about 6 kbp upstream to 11 kbp downstream of the TSS.



Identification of direct regulatory targets based on correlation

Next, we will identify the potential Tbx20 target genes whose down-regulated transcripts have a positional correlation with a Tbx20 peak in the range defined by the -6kbp/+11kbp plateau above. Correlations between the elements in the two data sets, as well as regions from the anchor (down-regulated transcripts) and partner (Tbx20 peaks) set, can be extracted based on the distances.

To retrieve the list of correlated transcripts and genes for the Tbx20 peaks in the -6kbp/+11kbp plateau, select the extraction of elements from the anchor set (the down-regulated transcripts), enter the range, and click on *Submit*.



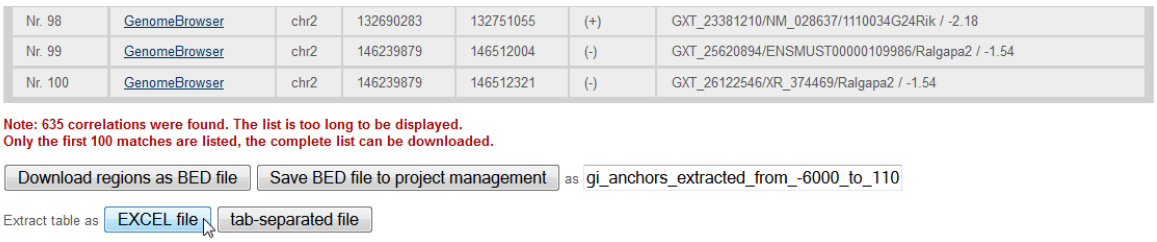
Al list of correlations will be shown, including distances and gene names (only the first 100 entries).

GenomeInspector: 635 correlations were found

Extracted Elements from Tbx20_ko_expression_transcripts_down.bed / Start with a correlation to Tbx20_peaks.bed within -6000 to 11000 bp

Number	GenomeBrowser	Chr.	Begin	End	Strand	Bed Id / Score
Nr. 1	GenomeBrowser	chr1	3999403	4409266	(-)	GXT_26095811/XM_006495473/Rp1 / -2.04
Nr. 2	GenomeBrowser	chr1	23995939	24005640	(-)	GXT_13127351/NM_026503/1110058L19Rik / -1.34
Nr. 3	GenomeBrowser	chr1	23995968	24005598	(-)	GXT_13007139/AK003789/1110058L19Rik / -1.48
Nr. 4	GenomeBrowser	chr1	24002966	24005630	(-)	GXT_24302361/ENSMUST00000155767/1110058L19Rik / -1.79
Nr. 5	GenomeBrowser	chr1	52845044	52885337	(+)	GXT_24324887/ENSMUST00000161125/Hibch / -1.69
Nr. 6	GenomeBrowser	chr1	52845046	52920860	(+)	GXT_13033472/NM_146108/Hibch / -1.56
Nr. 7	GenomeBrowser	chr1	52845048	52920860	(+)	GXT_12942462/AK076038/Hibch / -1.56
Nr. 8	GenomeBrowser	chr1	75383556	75384975	(+)	GXT_24322105/ENSMUST00000146705/Speg / -1.51
Nr. 9	GenomeBrowser	chr1	75384700	75387948	(+)	GXT_24322106/ENSMUST00000125118/Speg / -1.39
Nr. 10	GenomeBrowser	chr1	75384828	75391923	(+)	GXT_23717585/ENSMUST00000132228/Speg / -1.32

Press the *EXCEL file* download button at the end of the list, and open the file in Excel.



Nr. 98	GenomeBrowser	chr2	132690283	132751055	(+)	GXT_23381210/NM_028637/1110034G24Rik / -2.18
Nr. 99	GenomeBrowser	chr2	146239879	146512004	(-)	GXT_25620894/ENSMUST00000109986/Ralgapa2 / -1.54
Nr. 100	GenomeBrowser	chr2	146239879	146512321	(-)	GXT_26122546/XR_374469/Ralgapa2 / -1.54

Note: 635 correlations were found. The list is too long to be displayed.
Only the first 100 matches are listed, the complete list can be downloaded.

Download regions as BED file | Save BED file to project management as gi_anchors_extracted_from_-6000_to_110

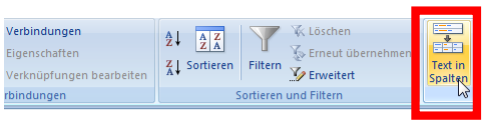
Extract table as EXCEL file | tab-separated file

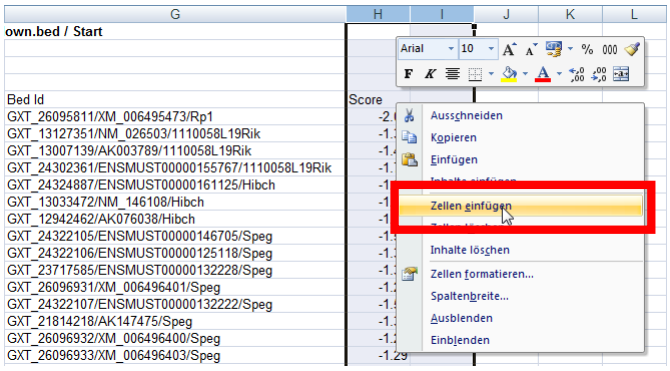
The *Bed Id* column for the anchor set contains the internal transcript identifiers (GXT_...), the transcript accession numbers, and the corresponding gene symbols. Note that you may need to adjust the column width to see the complete content.

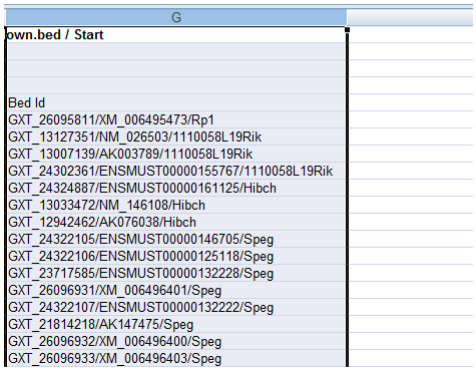
Extracted Elements from Tbx20_ko_expression_transcripts_down.bed / Start
with a correlation to Tbx20_peaks.bed
within -6000 to 11000 bp

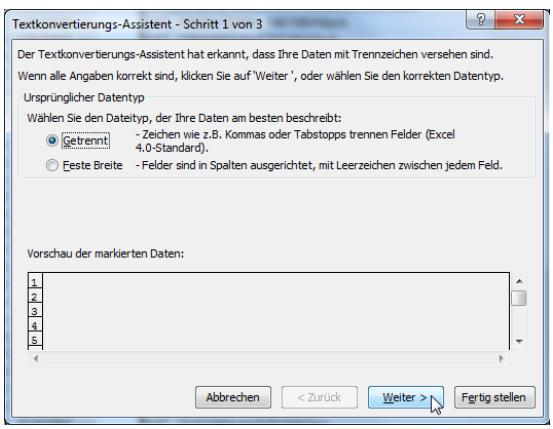
Number	GenomeBr	Chr.	Begin	End	Strand	Bed Id	Score
1	/cgi-bin//el	chr1	3999403	4409266	(-)	GXT_26095811/XM_006495473/Rp1	-2.04
2	/cgi-bin//el	chr1	23995939	24005640	(-)	GXT_13127351/NM_026503/1110058L19Rik	-1.34
3	/cgi-bin//el	chr1	23995968	24005598	(-)	GXT_13007139/AK003789/1110058L19Rik	-1.48
4	/cgi-bin//el	chr1	24002966	24005630	(-)	GXT_24302361/ENSMUST00000155767/1110058L19Rik	-1.79
5	/cgi-bin//el	chr1	52845044	52885337	(+)	GXT_24324887/ENSMUST00000161125/Hibch	-1.69
6	/cgi-bin//el	chr1	52845046	52920860	(+)	GXT_13033472/NM_146108/Hibch	-1.56
7	/cgi-bin//el	chr1	52845048	52920860	(+)	GXT_12942462/AK076038/Hibch	-1.56
8	/cgi-bin//el	chr1	75383556	75384975	(+)	GXT_24322105/ENSMUST00000146705/Speg	-1.51
9	/cgi-bin//el	chr1	75384700	75387948	(+)	GXT_24322106/ENSMUST00000125118/Speg	-1.39
10	/cgi-bin//el	chr1	75384828	75391923	(+)	GXT_23717585/ENSMUST00000132228/Speg	-1.32
11	/cgi-bin//el	chr1	75385158	75432320	(+)	GXT_26096931/XM_006496401/Speg	-1.29
12	/cgi-bin//el	chr1	75385512	75389104	(+)	GXT_24322107/ENSMUST00000132222/Speg	-1.51
13	/cgi-bin//el	chr1	75385610	75432304	(+)	GXT_21814218/AK147475/Speg	-1.31
14	/cgi-bin//el	chr1	75385676	75432320	(+)	GXT_26096932/XM_006496400/Speg	-1.29
15	/cgi-bin//el	chr1	75398588	75432320	(+)	GXT_26096933/XM_006496403/Speg	-1.29

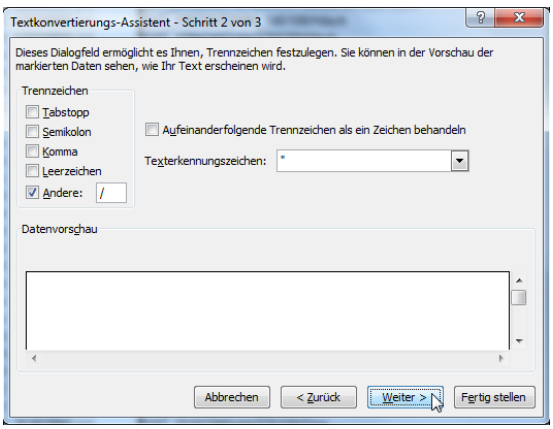
Use Excel functions to write the contents into separate columns: add two empty columns left of the Score column; then separate the text in the Bed Id column into different columns, using the slash (/) as separator.









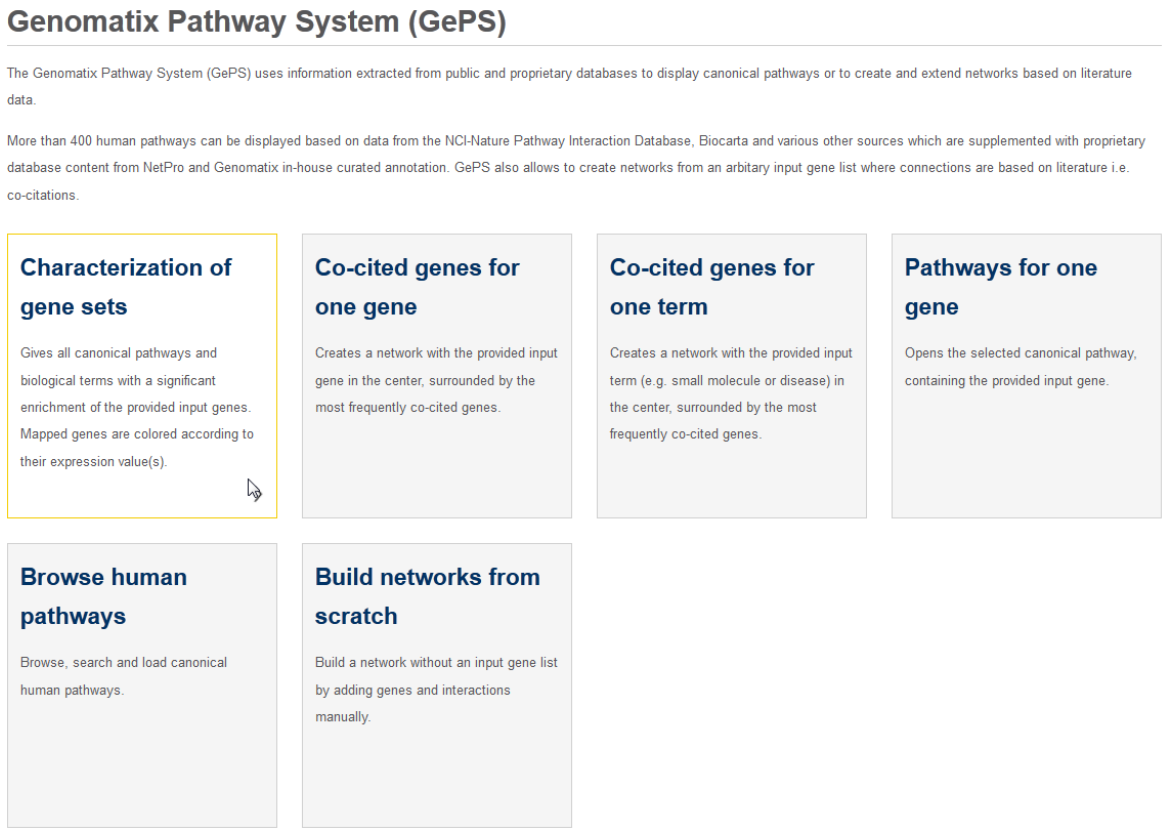


You should end up with a structure like this:

Extracted Elements from Tbx20_ko_expression_transcripts_down.bed / Start
with a correlation to Tbx20_peaks.bed
within -6000 to 11000 bp

Number	GenomeBr	Chr.	Begin	End	Strand	Bed Id			Score
1	/cgi-bin//el	chr1	3999403	4409266	(-)	GXT_26095811	XM_006495473	Rp1	-2.04
2	/cgi-bin//el	chr1	23995939	24005640	(-)	GXT_13127351	NM_026503	1110058L19Rik	-1.34
3	/cgi-bin//el	chr1	23995968	24005598	(-)	GXT_13007139	AK003789	1110058L19Rik	-1.48
4	/cgi-bin//el	chr1	24002966	24005630	(-)	GXT_24302361	ENSMUST00000155767	1110058L19Rik	-1.79
5	/cgi-bin//el	chr1	52845044	52885337	(+)	GXT_24324887	ENSMUST00000161125	Hibch	-1.69
6	/cgi-bin//el	chr1	52845046	52920860	(+)	GXT_13033472	NM_146108	Hibch	-1.56
7	/cgi-bin//el	chr1	52845048	52920860	(+)	GXT_12942462	AK076038	Hibch	-1.56
8	/cgi-bin//el	chr1	75383556	75384975	(+)	GXT_24322105	ENSMUST00000146705	Speg	-1.51
9	/cgi-bin//el	chr1	75384700	75387948	(+)	GXT_24322106	ENSMUST00000125118	Speg	-1.39
10	/cgi-bin//el	chr1	75384828	75391923	(+)	GXT_23717585	ENSMUST00000132228	Speg	-1.32
11	/cgi-bin//el	chr1	75385158	75432320	(+)	GXT_26096931	XM_006496401	Speg	-1.29
12	/cgi-bin//el	chr1	75385512	75389104	(+)	GXT_24322107	ENSMUST00000132222	Speg	-1.51
13	/cgi-bin//el	chr1	75385610	75432304	(+)	GXT_21814218	AK147475	Speg	-1.31
14	/cgi-bin//el	chr1	75385676	75432320	(+)	GXT_26096932	XM_006496400	Speg	-1.29
15	/cgi-bin//el	chr1	75398588	75432320	(+)	GXT_26096933	XM_006496403	Speg	-1.29

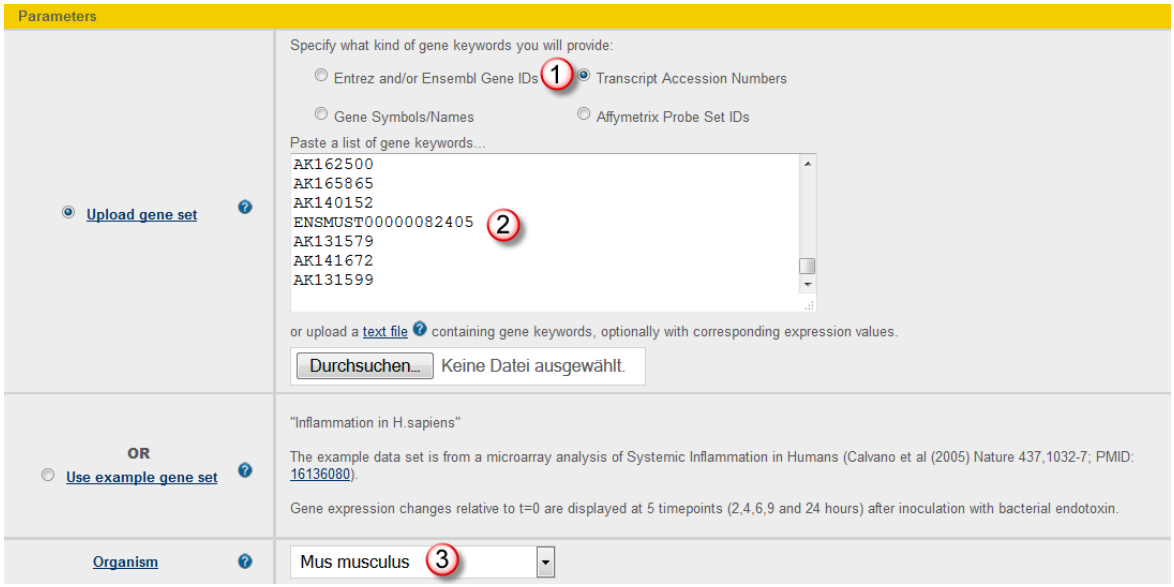
Open the Genomatix Pathway System from the navigation bar, and start a gene set characterization.



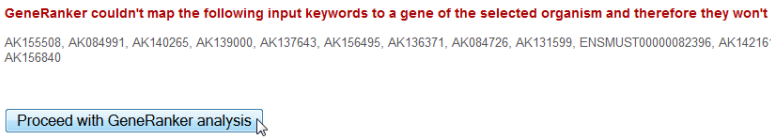
Extracted Elements from Tbx20_ko_expression_transcripts_down.bed / Start
with a correlation to Tbx20_peaks.bed
within -6000 to 11000 bp

Number	GenomeBr	Chr.	Begin	End	Strand	Bed Id			Score
1	/cgi-bin//el	chr1	3999403	4409266	(-)	GXT_26095811	XM_006495473	Rp1	-2.04
2	/cgi-bin//el	chr1	23995939	24005640	(-)	GXT_13127351	NM_026503	1110058L19Rik	-1.34
3	/cgi-bin//el	chr1	23995968	24005598	(-)	GXT_13007139	AK003789	1110058L19Rik	-1.48
4	/cgi-bin//el	chr1	24002966	24005630	(-)	GXT_24302361	ENSMUST00000155767	1110058L19Rik	-1.79
5	/cgi-bin//el	chr1	52845044	52885337	(+)	GXT_24324887	ENSMUST00000161125	Hibch	-1.69
6	/cgi-bin//el	chr1	52845046	52920860	(+)	GXT_13033472	NM_146108	Hibch	-1.56
7	/cgi-bin//el	chr1	52845048	52920860	(+)	GXT_12942462	AK076038	Hibch	-1.56
8	/cgi-bin//el	chr1	75383556	75384975	(+)	GXT_24322105	ENSMUST00000146705	Speg	-1.51
9	/cgi-bin//el	chr1	75384700	75387948	(+)	GXT_24322106	ENSMUST00000125118	Speg	-1.39
10	/cgi-bin//el	chr1	75384828	75391923	(+)	GXT_23717585	ENSMUST00000132228	Speg	-1.32
11	/cgi-bin//el	chr1	75385158	75432320	(+)	GXT_26096931	XM_006496401	Speg	-1.29
12	/cgi-bin//el	chr1	75385512	75389104	(+)	GXT_24322107	ENSMUST00000132222	Speg	-1.51
13	/cgi-bin//el	chr1	75385610	75432304	(+)	GXT_21814218	AK147475	Speg	-1.31
14	/cgi-bin//el	chr1	75385676	75432320	(+)	GXT_26096932	XM_006496400	Speg	-1.29
15	/cgi-bin//el	chr1	75398588	75432320	(+)	GXT_26096933	XM_006496403	Speg	-1.29

Copy the transcript accession numbers from the Excel list to the gene keyword input field (duplicates will be removed by the system). Select *Transcript Accession Numbers* as the keyword type. Select *Mus musculus* as organism, and start the search.

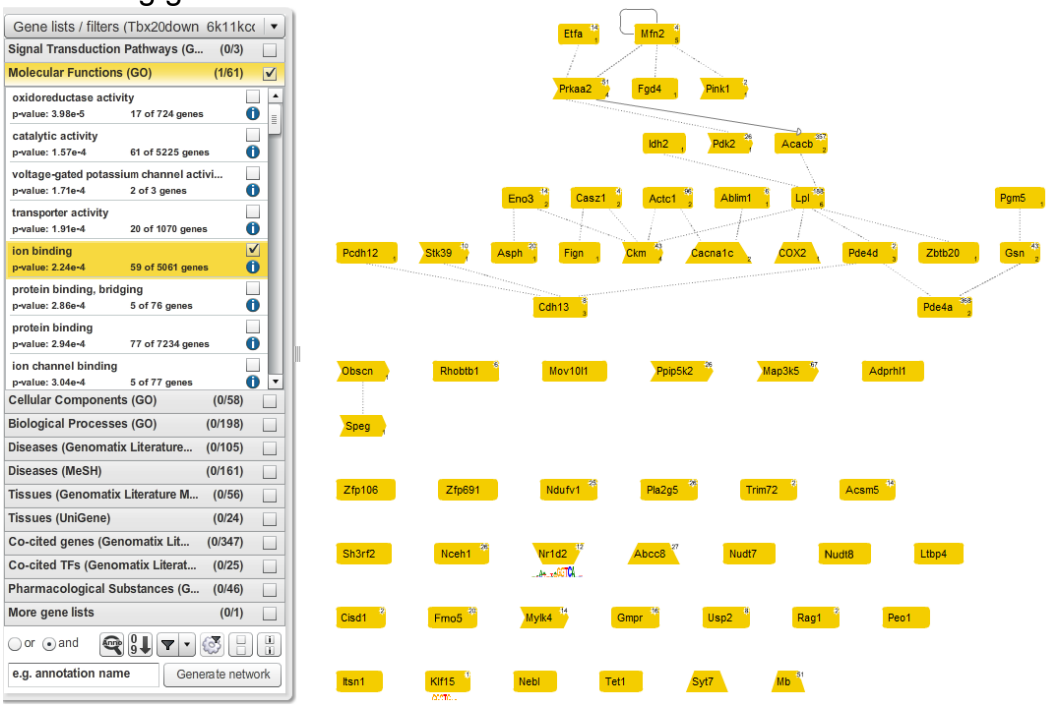


Some accession numbers will not be mapped to a gene ID; ignore the warning the program gives you, and proceed with the analysis.

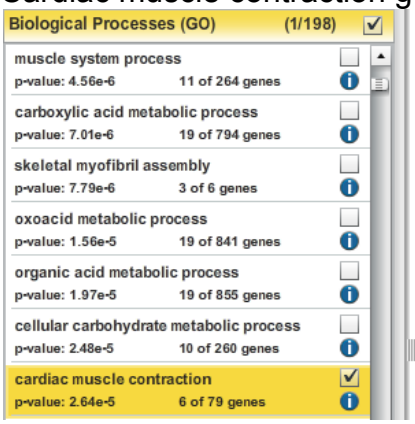


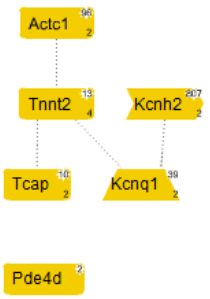
A total of 190 genes are found in this way. Overrepresented terms include *ion binding* in *GO: Molecular Function*, *cardiac muscle contraction* in *GO: Biological Processes*, and *cardiomyopathies* in *Diseases*. The pathway graphs below use the hierarchical layout, which you can activate with the leftmost *Layout* button in the lower control bar:

Ion binding gene network:

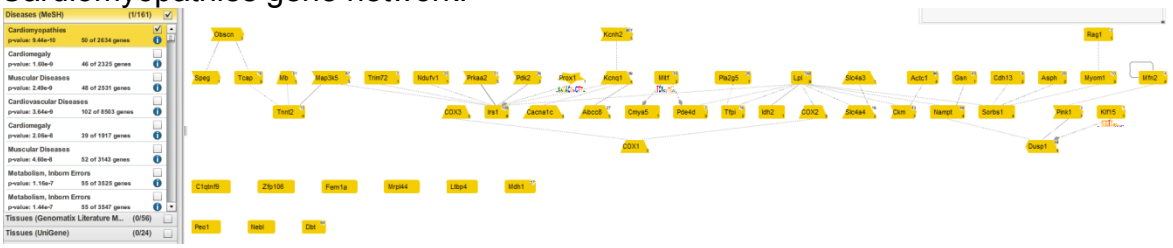


Cardiac muscle contraction gene network:

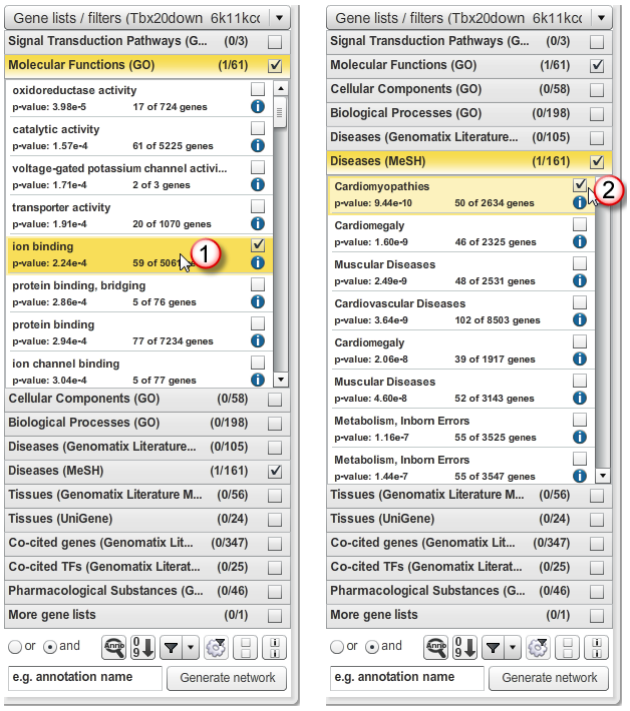




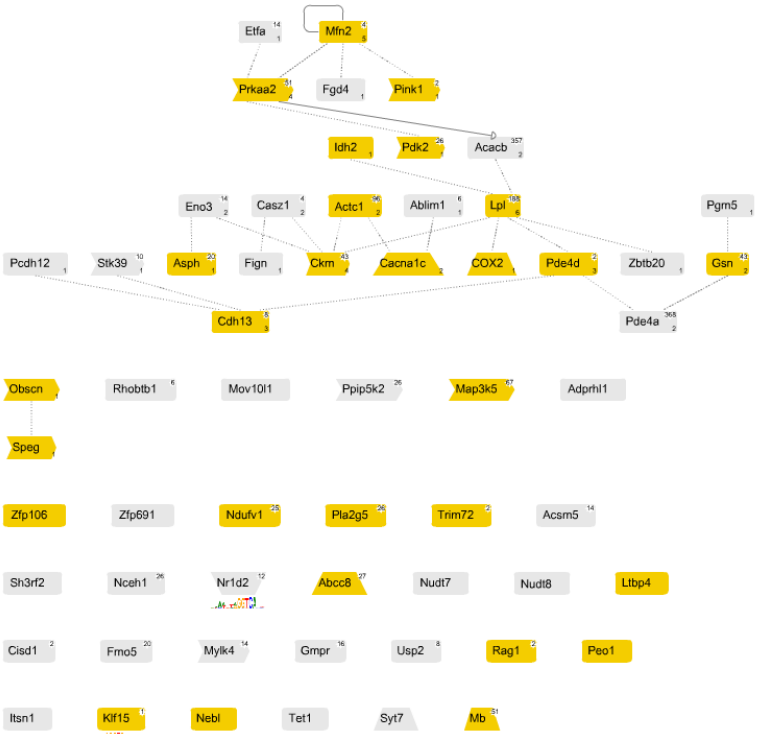
Cardiomyopathies gene network:



Many of the ion binding genes are also associated with cardiomyopathies, as can be seen by selecting the ion binding network, and then ticking the checkbox for the cardiomyopathies associated genes:



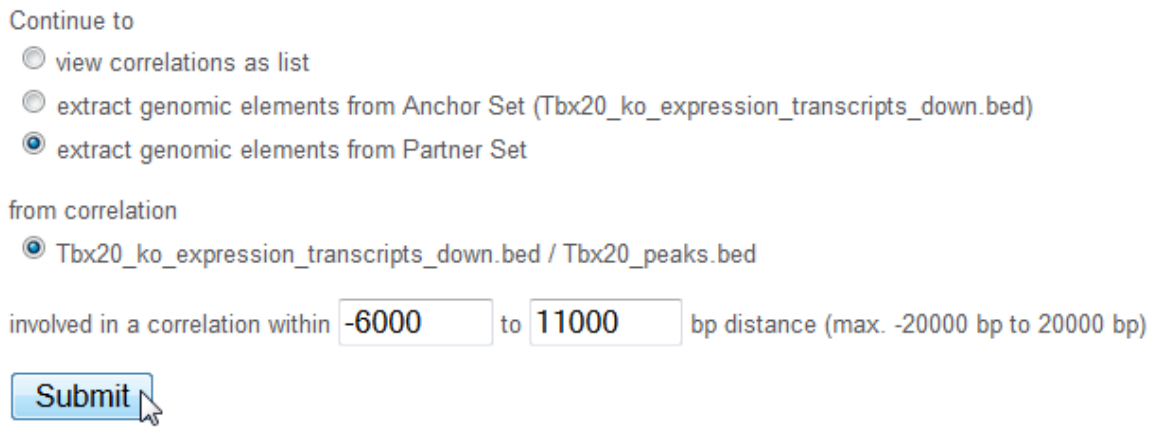
Only ion binding network genes fulfilling both criteria are shown with a colored background.



In-depth transcription factor binding site analysis of correlated peaks

The next analysis step will take a closer look at the peak regions which form the correlation plateau in the GenomeInspector graph. You will retrieve a BED file of the correlated peaks, prepare it in the BED file toolbox for downstream analysis, and find common transcription factor binding site patterns that include Tbx20 binding sites, which will then be assessed further.

Go back to the GenomeInspector output or open the result from the project management, and select the extraction of elements from the partner set (the Tbx20 peaks), again setting the distance range to -6kbp/+11kbp.



208 correlated peaks are found.

GenomeInspector: 208 correlations were found

Extracted Elements from Tbx20_peaks.bed
with a correlation to Tbx20_ko_expression_transcripts_down.bed / Start
within -6000 to 11000 bp

Number	GenomeBrowser	Chr.	Begin	End	Strand	Bed Id / Score
Nr. 1	GenomeBrowser	chr1	4412567	4412753	(+)	1 / 9.12e-13
Nr. 2	GenomeBrowser	chr1	24010974	24011102	(+)	12 / 1.01e-06
Nr. 3	GenomeBrowser	chr1	52844928	52845037	(+)	40 / 1.66e-07
Nr. 4	GenomeBrowser	chr1	75393329	75393513	(+)	70 / 1.31e-10
Nr. 5	GenomeBrowser	chr1	75549351	75549552	(+)	71 / 1.47e-14
Nr. 6	GenomeBrowser	chr1	79776017	79776139	(+)	77 / 2.95e-05
Nr. 7	GenomeBrowser	chr1	82291491	82291596	(+)	78 / 3.87e-06
Nr. 8	GenomeBrowser	chr1	97761621	97761790	(+)	93 / 8.83e-09
Nr. 9	GenomeBrowser	chr1	118479426	118479585	(+)	108 / 2.31e-10
Nr. 10	GenomeBrowser	chr1	118481924	118482088	(+)	109 / 1.1e-08
Nr. 11	GenomeBrowser	chr1	135850907	135851029	(+)	137 / 1.16e-05
Nr. 12	GenomeBrowser	chr1	155234365	155234481	(+)	157 / 7.66e-07

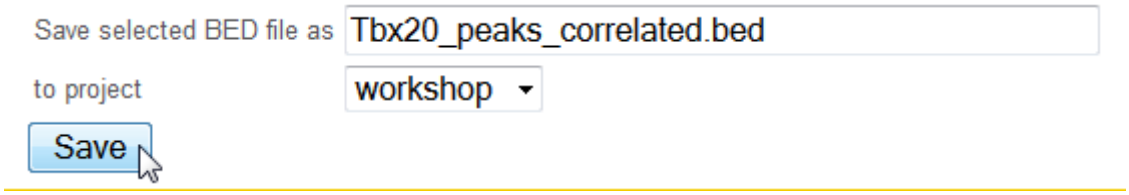
Scroll down to the end of the list and save the regions as BED file in your project management.

Nr. 100	GenomeBrowser	chr8	68908384	68908503	(+)	2793 / 2.95e-05

Note: 208 correlations were found. The list is too long to be displayed.
Only the first 100 matches are listed, the complete list can be downloaded.

Download regions as BED file | Save BED file to project management as Tbx20_peaks_correlated.bed

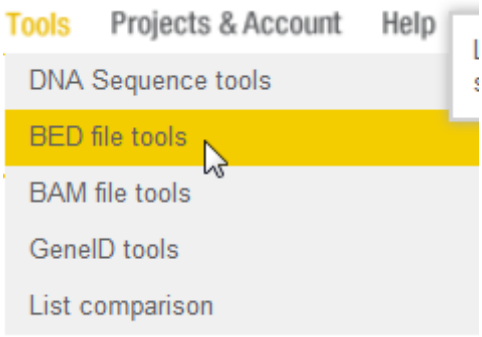
Extract table as EXCEL file | tab-separated file



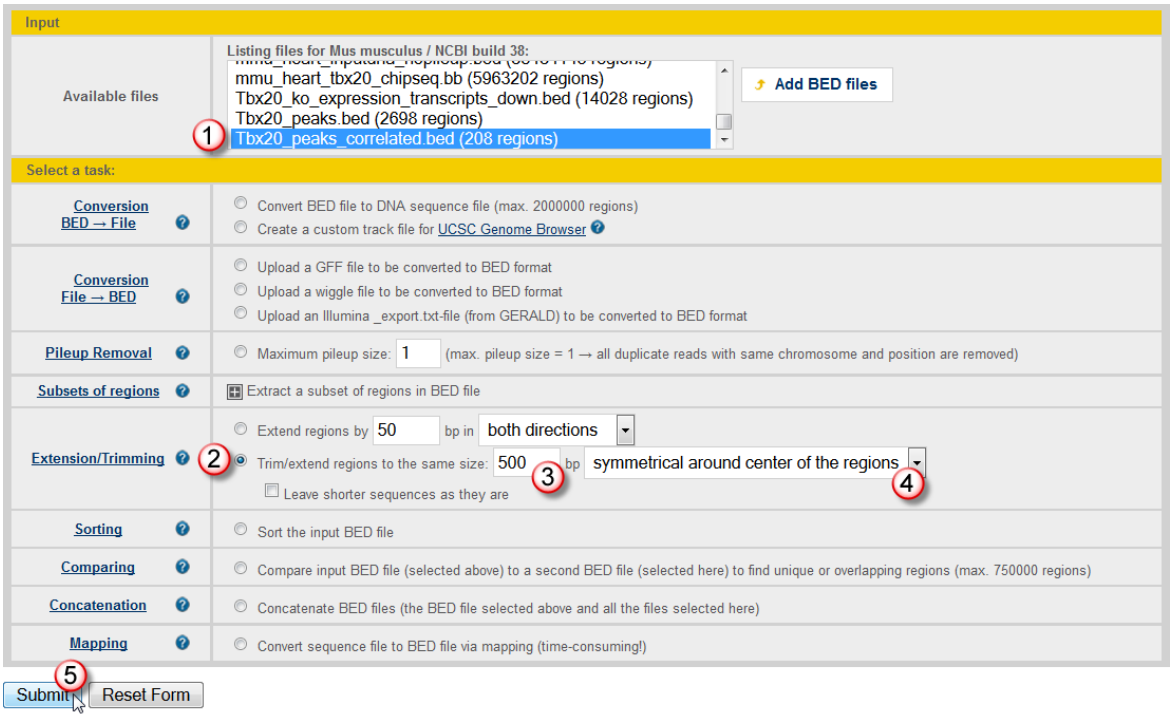
Trimming and conversion to sequence

For downstream analysis, the peak data need to be modified in two ways: one, some of the regions are too long to be accepted as input for the FrameWorker program, which we will use for detection of common transcription factor binding site patterns in the peaks. Therefore, we will give the peaks a uniform length. Secondly, FrameWorker needs sequences as input; therefore we will generate a sequence file from the modified BED file.

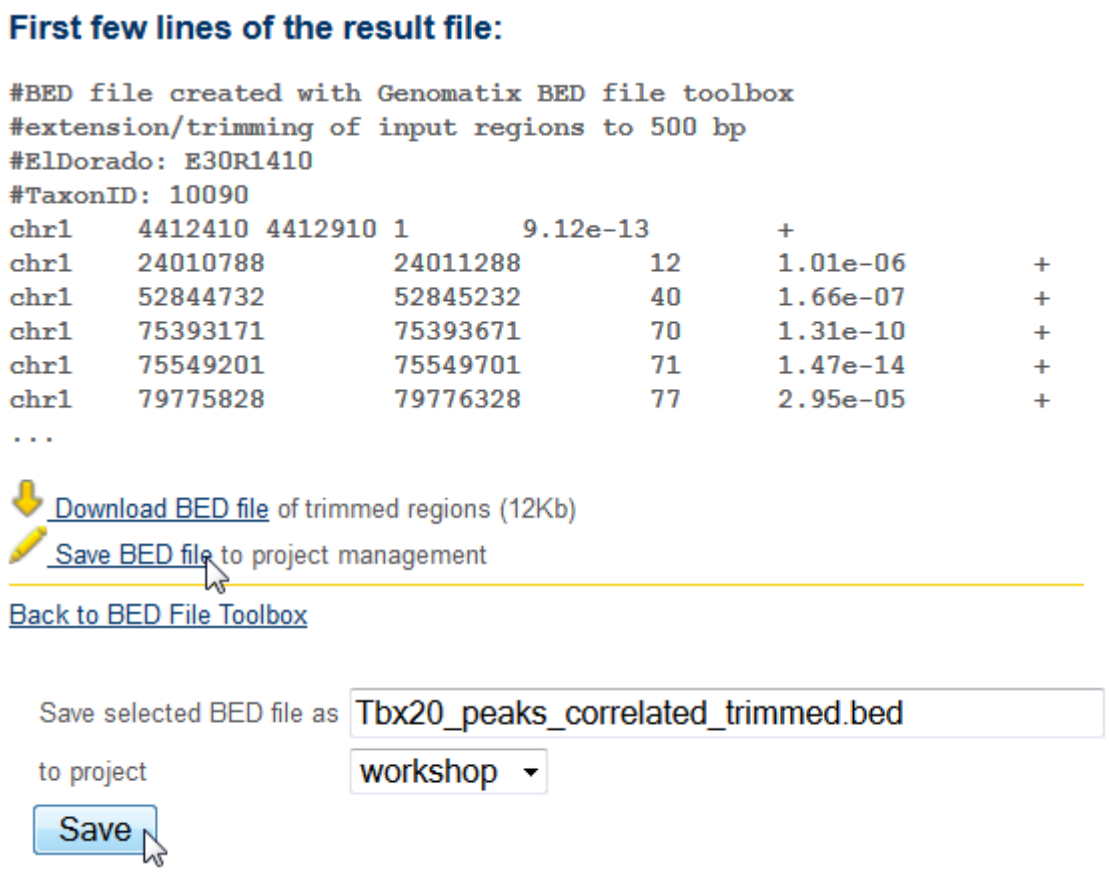
Open the BED file tools in the navigation bar.



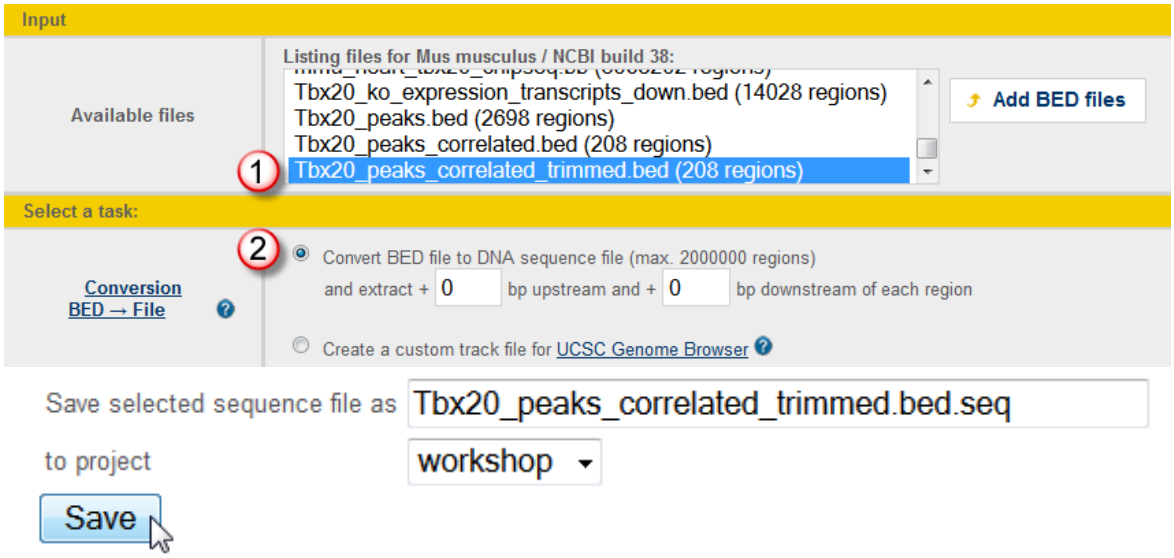
Select the file with the correlated peaks in the list. Then select the option *Trim regions to the same size*, set the size to 500bp, and select the option *symmetrical around the center of the regions*; then press *Submit*.



Save the file to the project management.



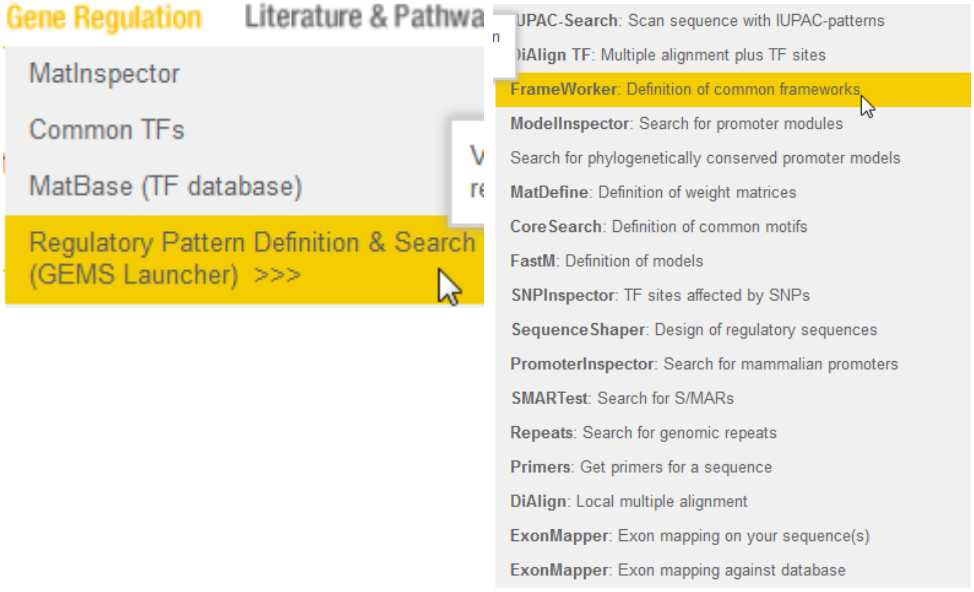
Open the BED file tools once more to convert the trimmed file to a sequence file. Select the trimmed file in the list, and activate the Convert BED file to DNA sequence file function. Start the conversion, and save the result in your project management.



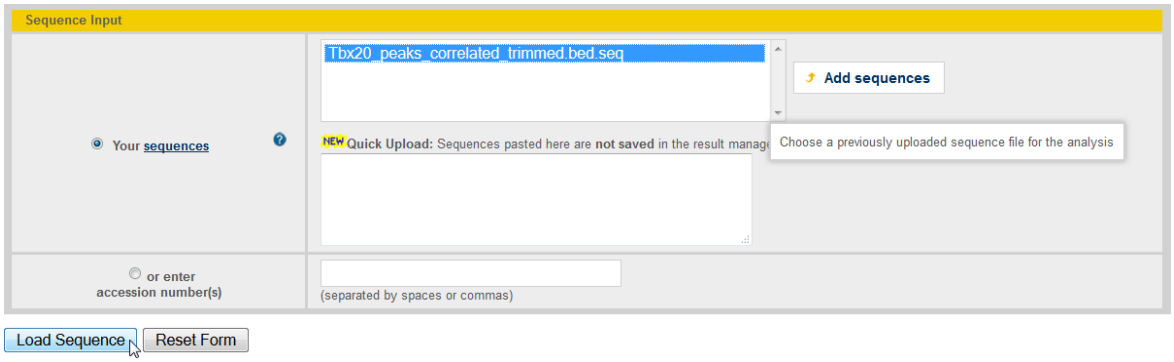
FrameWorker: common TFBS patterns

The next step in the analysis will employ FrameWorker, which searches for common patterns of TF binding sites in a set of input sequences – here, the Tbx20 peaks which are correlated with genes that are down-regulated in Tbx20 knock-out mouse hearts.

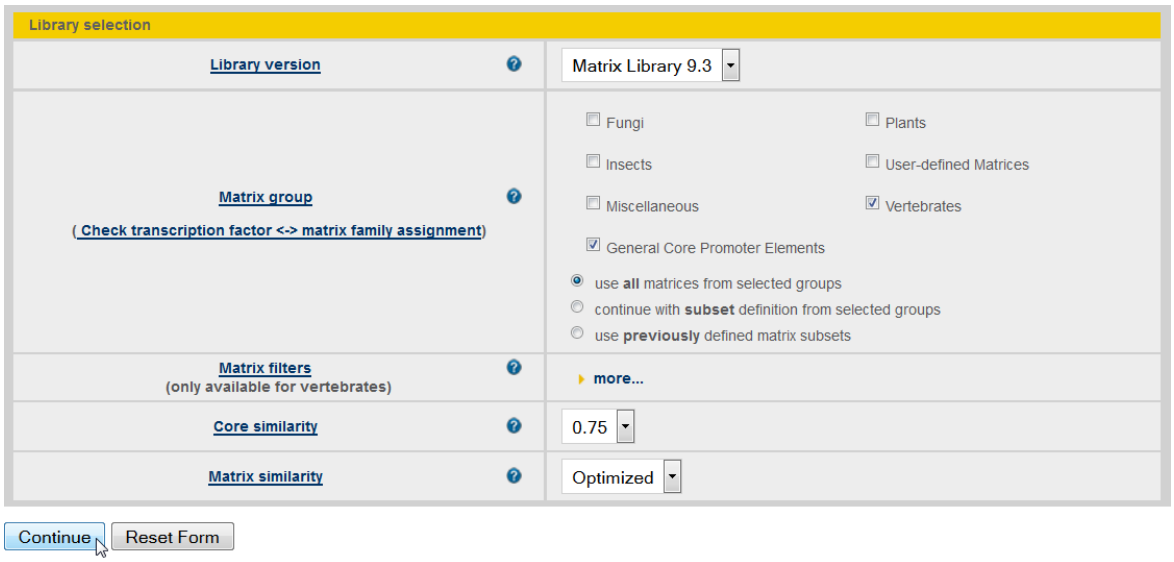
You'll find the program in the navigation bar under *Gene Regulation - Regulatory Pattern Definition & Search - FrameWorker*:



Select the saved sequence file in the list and continue by clicking the *Load Sequence* button.



The next step lets you choose the elements the frameworks can be made up of. For this analysis, please leave the settings at the defaults.

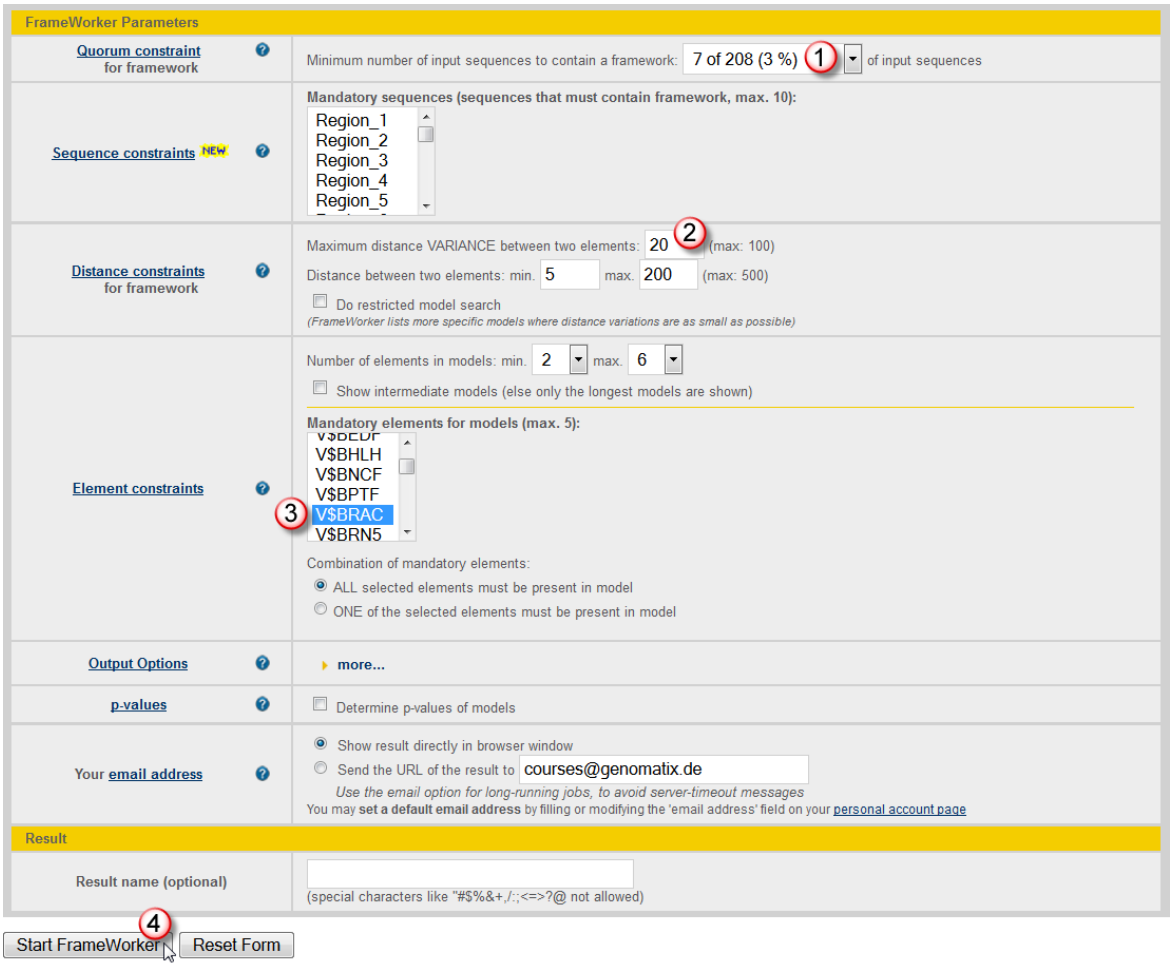


Ignore the warning on the next page, and press *Continue*.

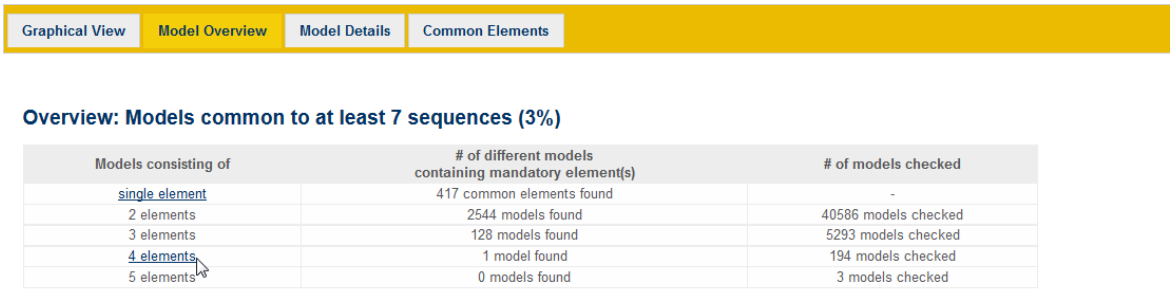
In the next step, parameters defining the stringency of the pattern search are set. Specifically the quorum constraint, and sometimes also the distance constraints, are usually changed in an iterative process, checking the result and adapting the stringency so that a handful of patterns of the desired complexity are found, which are then further evaluated.

For this example, please set the parameters as follows:
- Quorum constraint = 7 of 208
- Distance constraints: maximum distance variance = 20
- Element constraints: V$BRAC (the binding site for Tbx20) mandatory element.

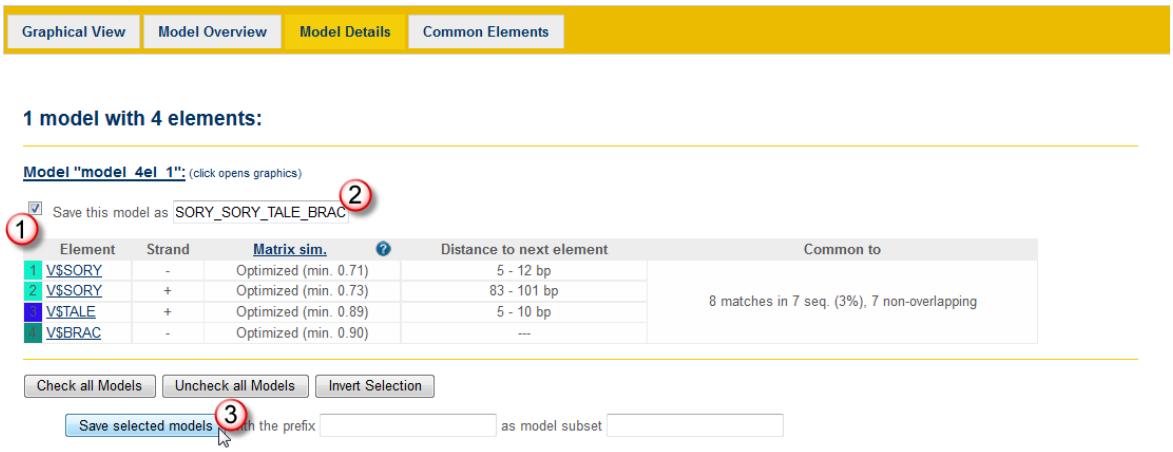
Then start the analysis.



One model with 4 elements is found. Click the link to jump to the description.



The model consists of two SORY sites, a TALE site, and the mandatory BRAC site. You can click on the links for each binding site symbol to find more information about it. The V$SORY binding site family binds, among others, the Sox6 protein, which has a role in cardiomyocyte differentiation. V$TALE can bind Meis1, which is a regulator of the cardiomyocyte cell cycle.



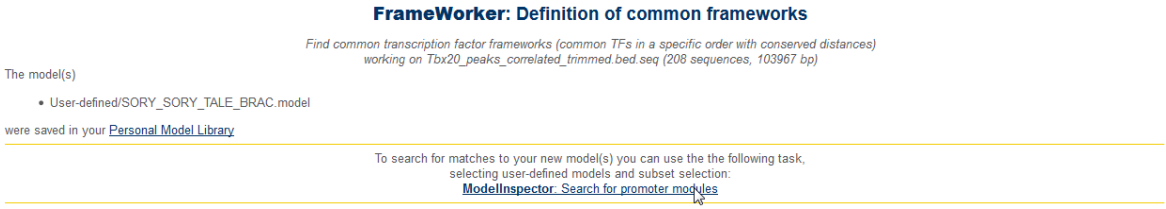
From the associated biology, the pattern could be of interest. Please tick its checkbox, marking it for saving, give it a name, e.g. SORY_SORY_TALE_BRAC, and press the *Save selected models* button.

ModelInspector: check for relevant biology

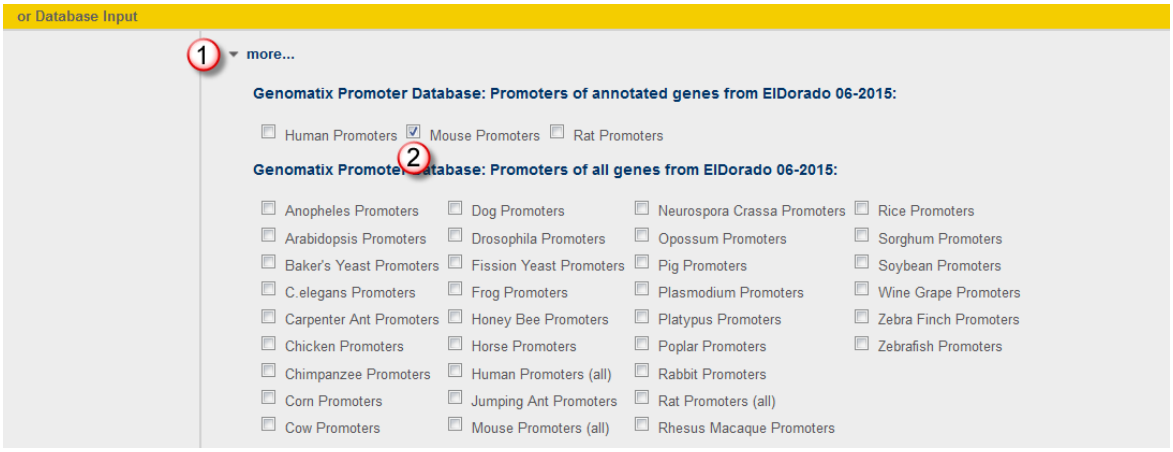
For further evaluation of the model, we will run a ModelInspector search, and try to find patterns matches in a mouse promoter database. The output will include an overrepresentation analysis for GO terms in the categories biological process, molecular function, and cellular component, which allows us to assess whether the binding site pattern is associated with relevant biology.

ModelInspector uses model definitions as are generated by FrameWorker to scan DNA sequences for matches. A model is defined as a set of various individual elements (here: transcription factor binding sites), their strand orientation, their sequential order, and their distance ranges.

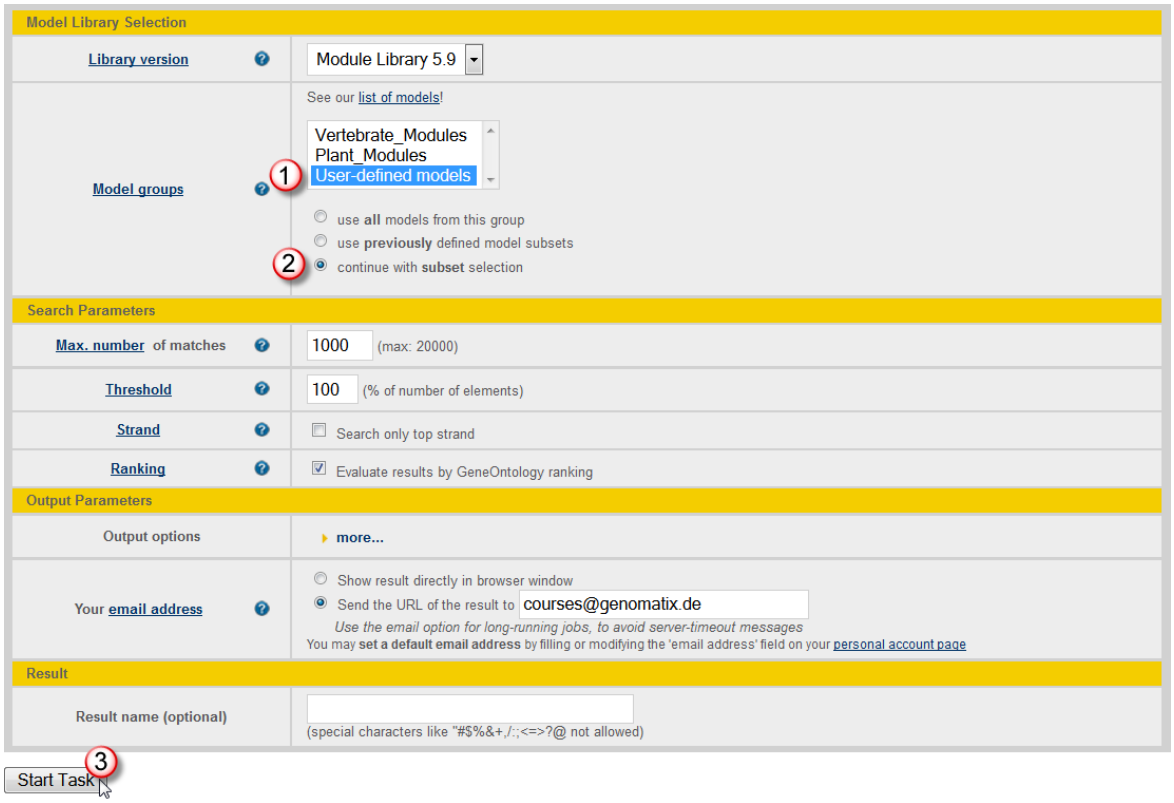
Click on the *ModelInspector* link on the notification page you see after the model has been saved.



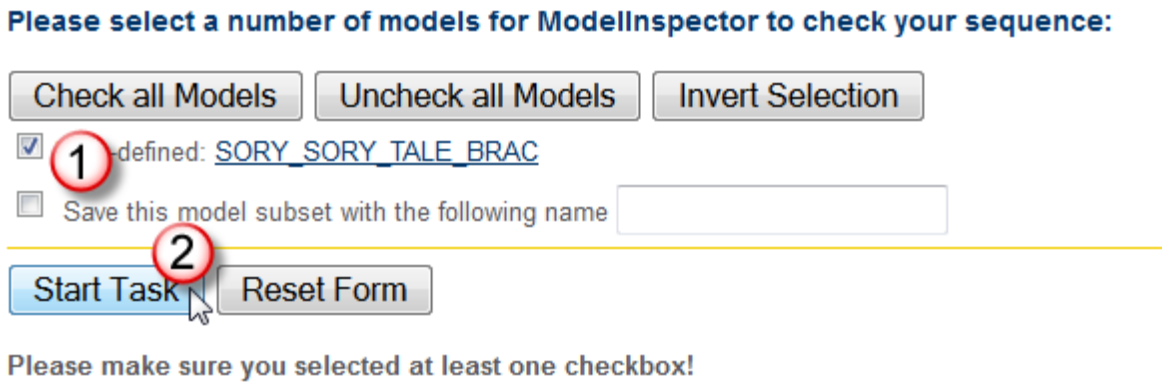
We will scan all mouse promoters of annotated genes. Click the *more...* option in the *Database Input* section to display the available parameters, then check *Mouse Promoters* in the section *Promoters of annotated genes* and proceed with *Load Sequence*.



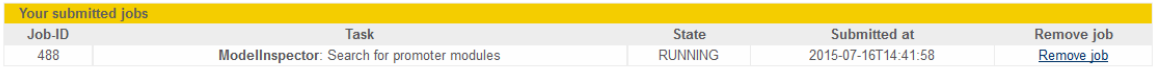
On the following parameter screen, select *User defined models* and *continue with subset selection* and continue.



Select the newly saved model in the list and start.

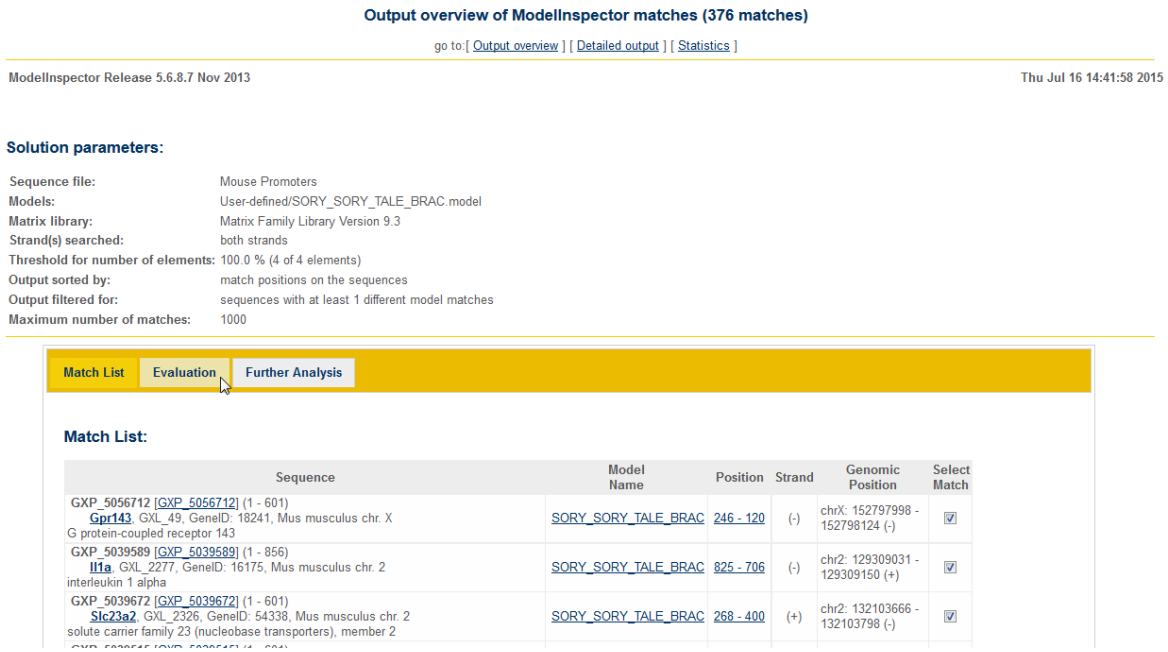


The analysis will run in the background; repeatedly check in your project management if the job is still running:



When it is done, click the link in the results directory to open the result page.

There are 376 matches in the promoter database. Click on *Evaluation* in the table header to display the GO statistics for the matching genes.



Output overview of ModelInspector matches (376 matches)

go to:[Output overview] [Detailed output] [Statistics]

ModelInspector Release 5.6.8.7 Nov 2013 — Thu Jul 16 14:41:58 2015

Solution parameters:

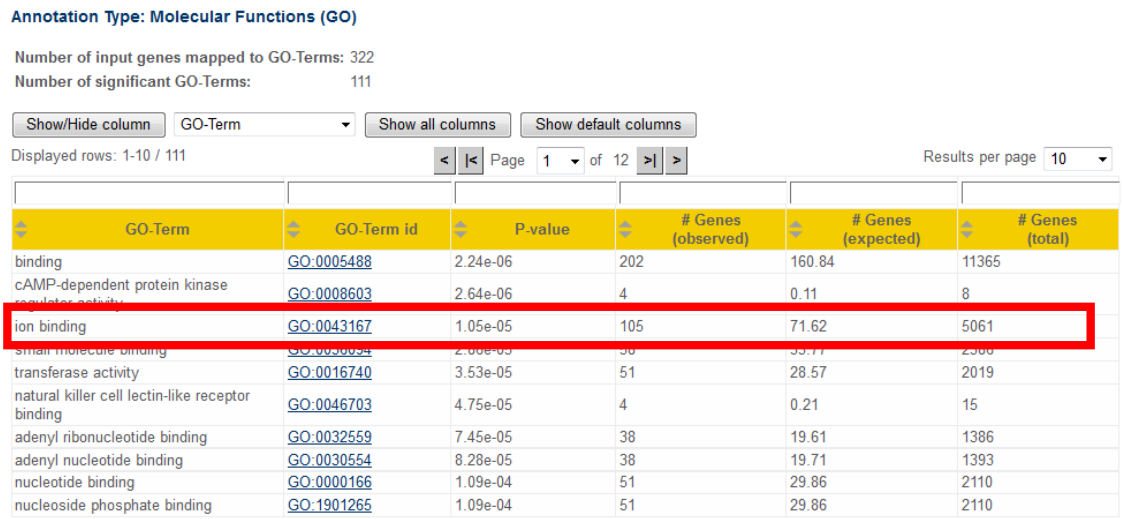
Sequence file:	Mouse Promoters
Models:	User-defined/SORY_SORY_TALE_BRAC.model
Matrix library:	Matrix Family Library Version 9.3
Strand(s) searched:	both strands
Threshold for number of elements:	100.0 % (4 of 4 elements)
Output sorted by:	match positions on the sequences
Output filtered for:	sequences with at least 1 different model matches
Maximum number of matches:	1000

Match List | Evaluation | Further Analysis

Match List:

Sequence	Model Name	Position	Strand	Genomic Position	Select Match
GXP_5056712 [GXP_5056712] (1 - 601) Gpr143, GXL_49, GeneID: 18241, Mus musculus chr. X G protein-coupled receptor 143	SORY_SORY_TALE_BRAC	246 - 120	(-)	chrX: 152797998 - 152798124 (-)	☑
GXP_5039589 [GXP_5039589] (1 - 856) Il1a, GXL_2277, GeneID: 16175, Mus musculus chr. 2 interleukin 1 alpha	SORY_SORY_TALE_BRAC	825 - 706	(-)	chr2: 129309031 - 129309150 (+)	☑
GXP_5039672 [GXP_5039672] (1 - 601) Slc23a2, GXL_2326, GeneID: 54338, Mus musculus chr. 2 solute carrier family 23 (nucleobase transporters), member 2	SORY_SORY_TALE_BRAC	268 - 400	(+)	chr2: 132103666 - 132103798 (-)	☑

One of the overrepresented terms in the *Molecular Functions* category is *ion binding*, with a model match in the promoters of 105 genes. The same term was also overrepresented in the list of genes that were downregulated in Tbx20 knockout hearts and in the subset of down-regulated transcripts with a correlated Tbx20 ChIP-Seq peak in a +/- 10kb window around the TSS. This last analysis shows that the model finds the term also in all promoters, independent of expression or ChIP-Seq analysis results.



Annotation Type: Molecular Functions (GO)

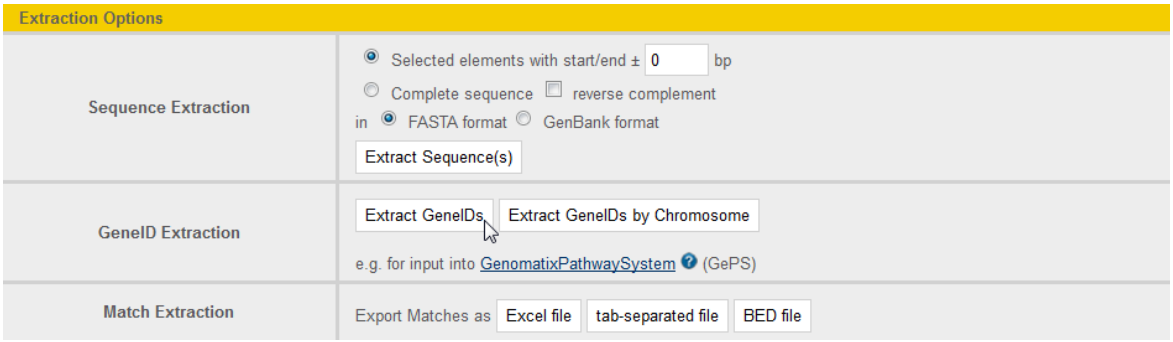
Number of input genes mapped to GO-Terms: 322
Number of significant GO-Terms: 111

Show/Hide column | GO-Term | Show all columns | Show default columns

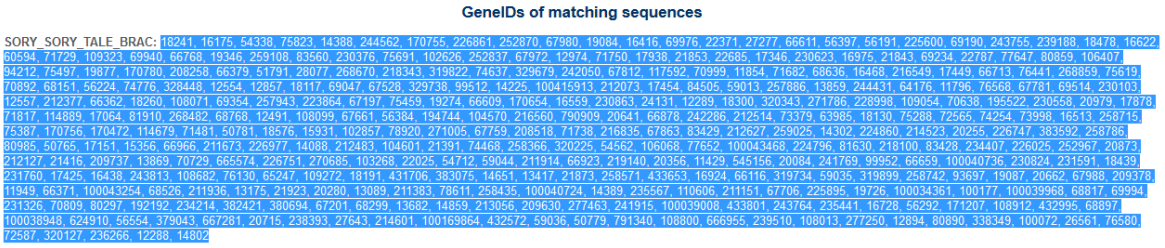
Displayed rows: 1-10 / 111 — < |< Page 1 of 12 >| > — Results per page 10

GO-Term	GO-Term id	P-value	# Genes (observed)	# Genes (expected)	# Genes (total)
binding	GO:0005488	2.24e-06	202	160.84	11365
cAMP-dependent protein kinase regulator activity	GO:0008603	2.64e-06	4	0.11	8
ion binding	GO:0043167	1.05e-05	105	71.62	5061
small molecule binding	GO:0036094	2.66e-05	56	33.77	2366
transferase activity	GO:0016740	3.53e-05	51	28.57	2019
natural killer cell lectin-like receptor binding	GO:0046703	4.75e-05	4	0.21	15
adenyl ribonucleotide binding	GO:0032559	7.45e-05	38	19.61	1386
adenyl nucleotide binding	GO:0030554	8.28e-05	38	19.71	1393
nucleotide binding	GO:0000166	1.09e-04	51	29.86	2110
nucleoside phosphate binding	GO:1901265	1.09e-04	51	29.86	2110

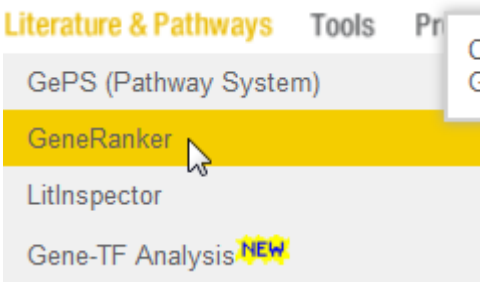
In order to find more associated biology for the genes with a model match in the promoter, we'll use the program GeneRanker. Go back to the ModelInspector match list output, and scroll down to the end of the page. Here, press the 'Extract GeneIDs' button to open a page showing the GeneIDs.



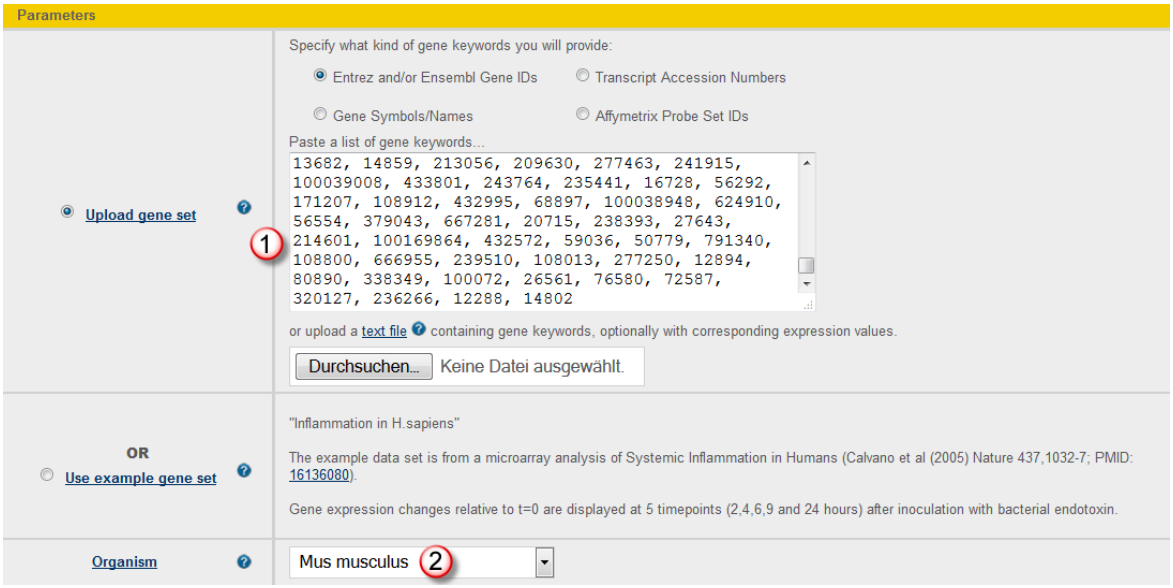
Select the Gene IDs on this page and copy them to the clipboard.



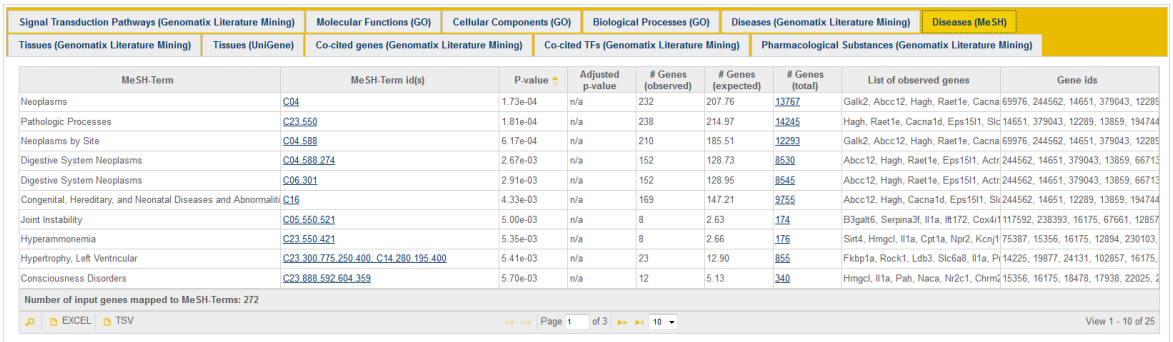
Open GeneRanker from the Literature & Pathways menu in the navigation bar.

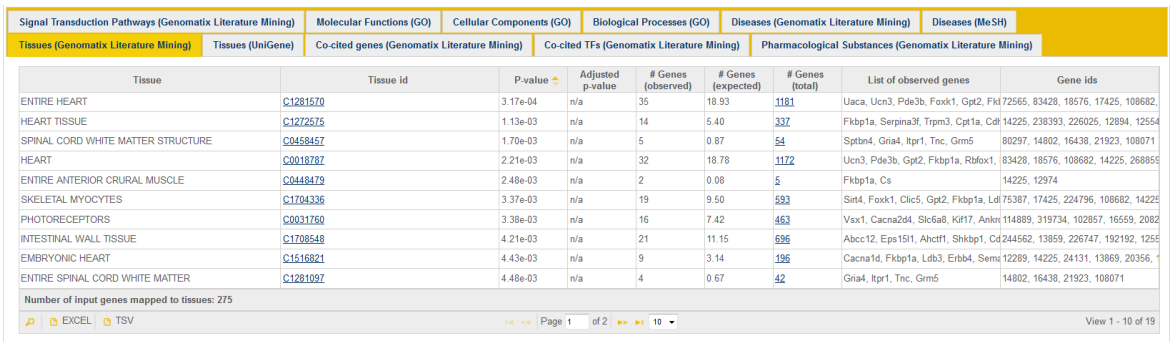


Paste the Gene IDs from the clipboard into the keyword field, select the mouse as organism, and start the analysis.



The result shows the cardiomyopathy *hypertrophy, left ventricular* in the top ten overrepresented *MeSH Disease* terms, and several heart associated terms in the category *Tissues (Genomatix Literature Mining)*.

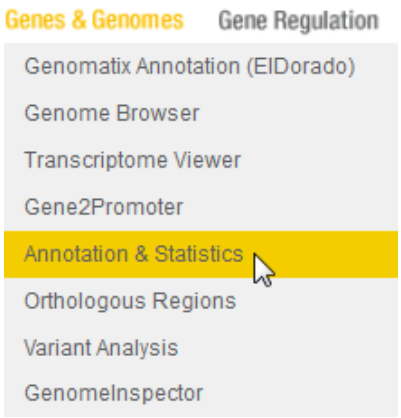




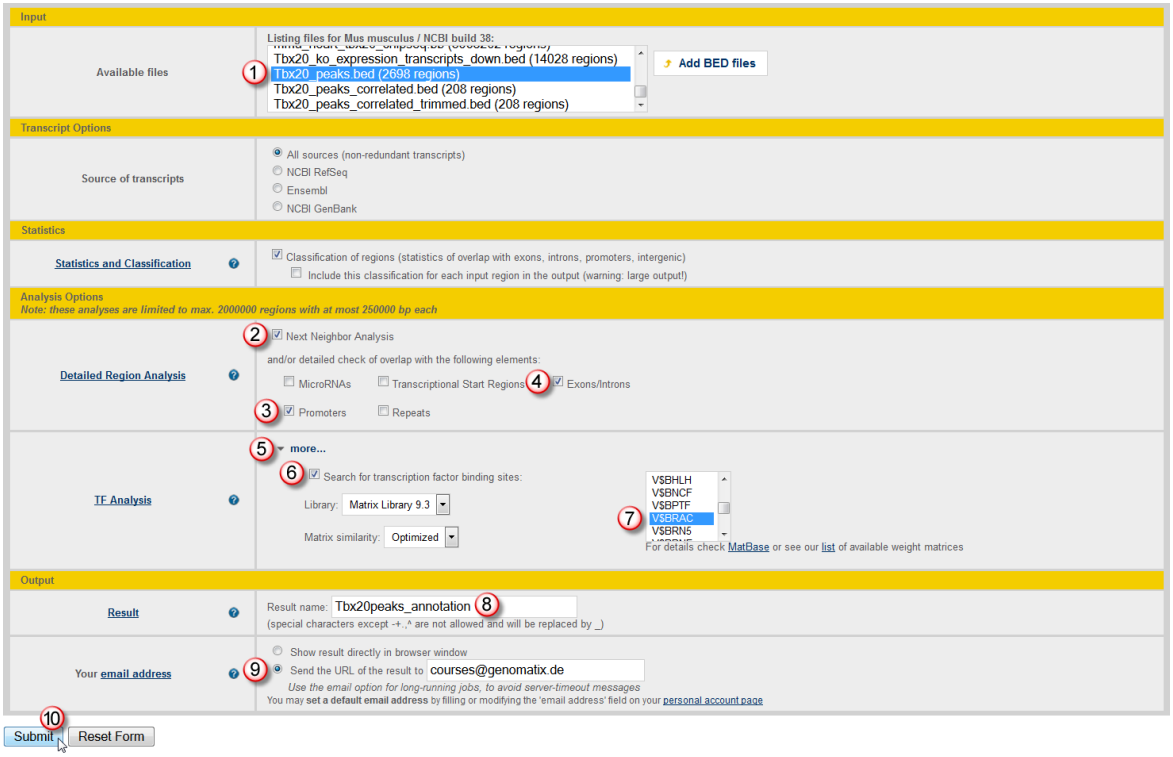
Annotation of Tbx20 binding regions – target prediction

An alternative way for finding potential regulatory targets of a transcription factor based on ChIP-Seq peaks, which can also be applied in the absence of expression data, is to analyze the genomic annotation in the vicinity of the TF peak positions and look for overlapping and neighboring promoters and gene loci.

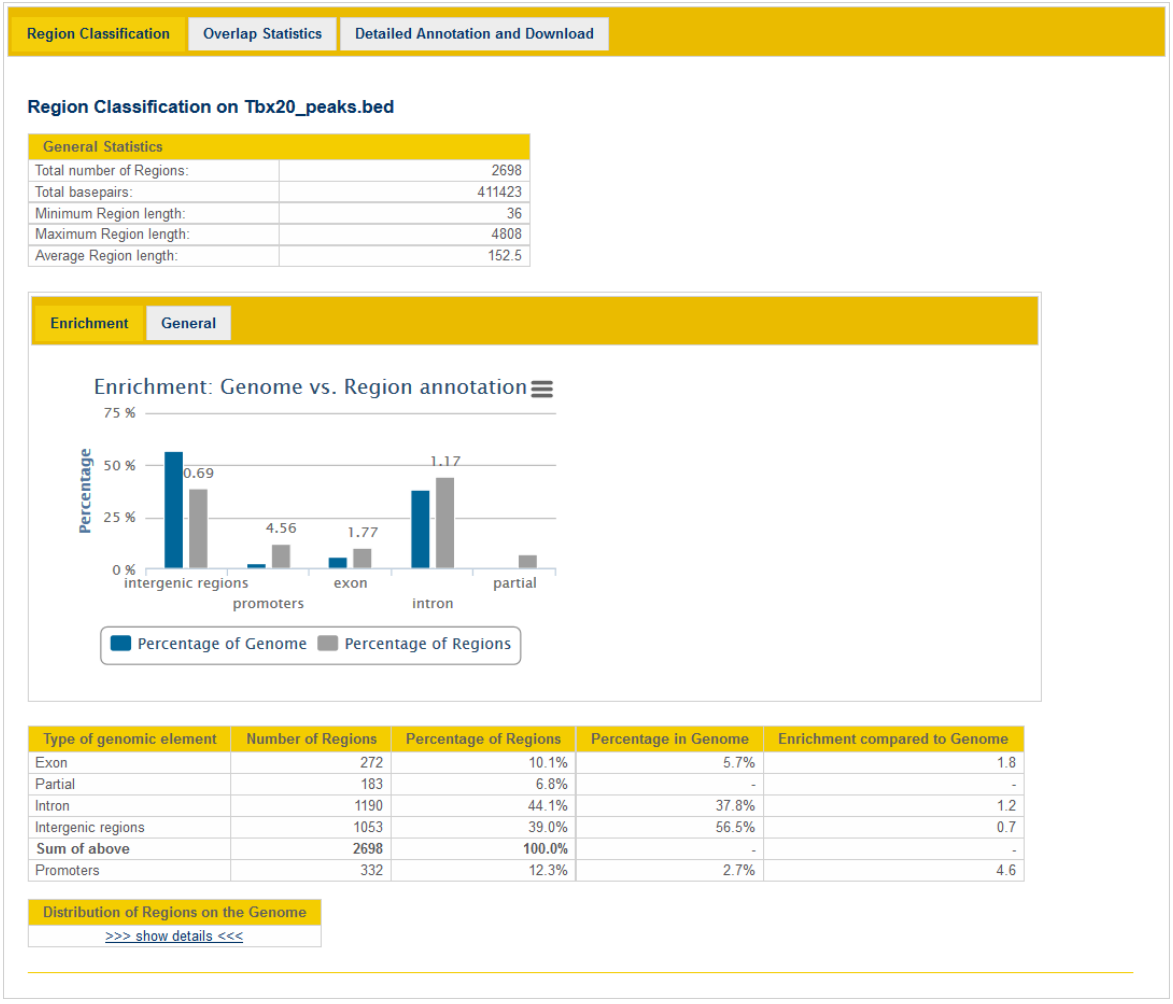
The program “Annotation & Statistics” annotates your input regions for features such as promoter overlaps or neighboring loci. Please start this task from the Genes & Genomes menu in the navigation bar:



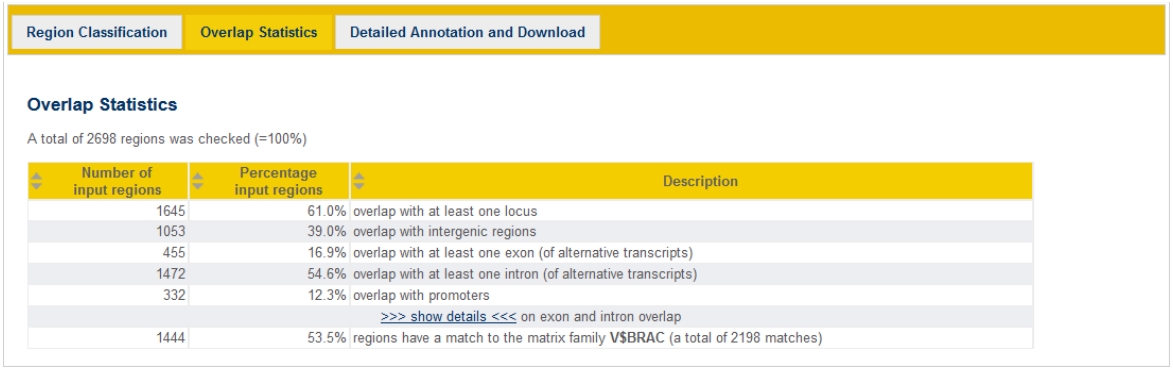
Please set the analysis parameters as below: select the BED file with the Tbx20 peaks from the BED file list, and activate the *Next Neighbor Analysis*, *Exons/Introns*, and *Promoters* checkboxes, This is necessary for identification of neighboring and overlapping promoters and loci. To include the information which peaks have a match for a Tbx20 binding site, click on the TF analysis *more...* option, tick the TFBS search checkbox, and select V$BRAC from the binding site list. Provide a result name, make sure that you selected the e-mail option, and start the analysis. As we have more than 2000 regions to analyze in detail, the analysis will take about 10 minutes.



When the analysis has completed, please open it in the project management. A classification table displays the numbers for the overlap of genome annotation with your input regions.

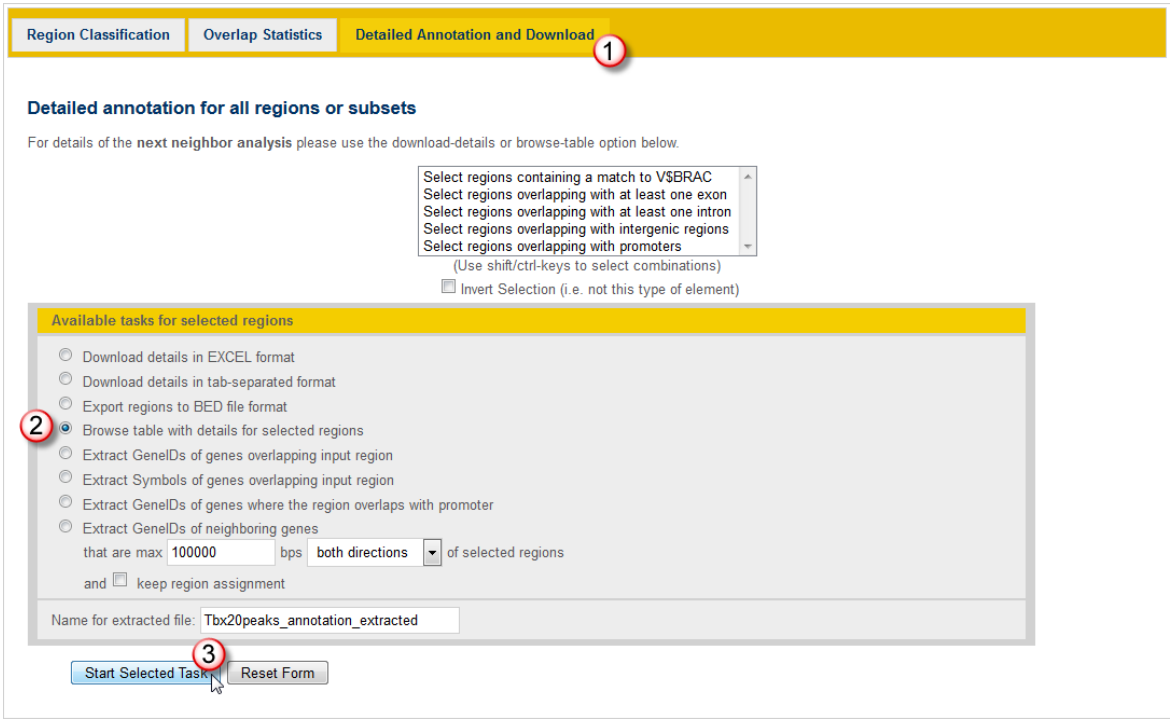


Overlap details can be viewed in the Overlap Statistics section.

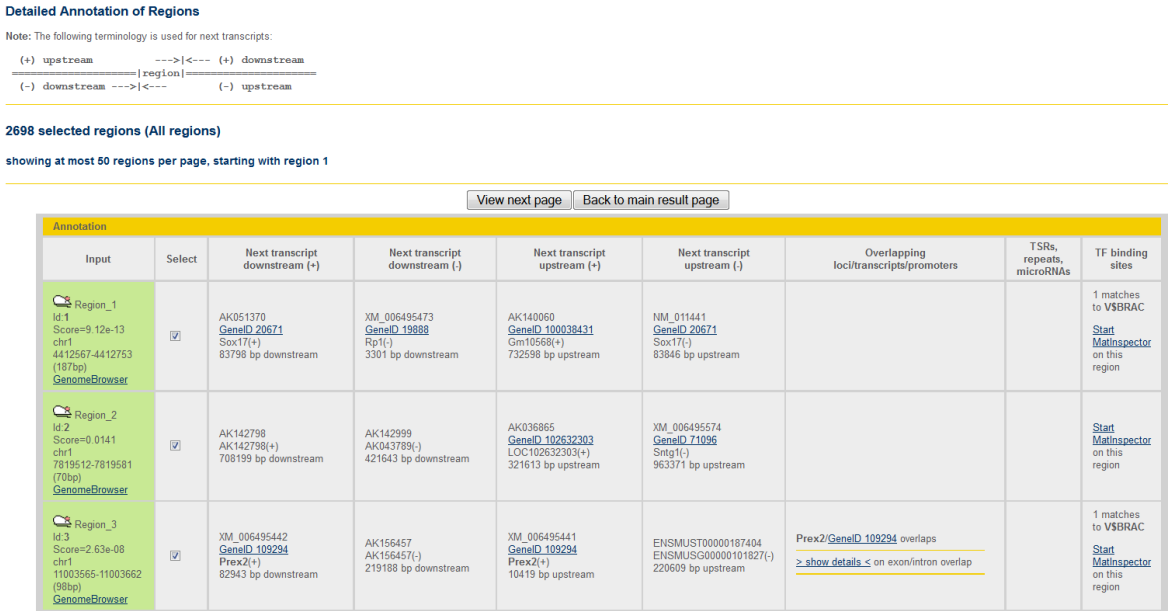


Based on this annotation, different data sets can be generated. Select *Detailed Annnotation and Download*.

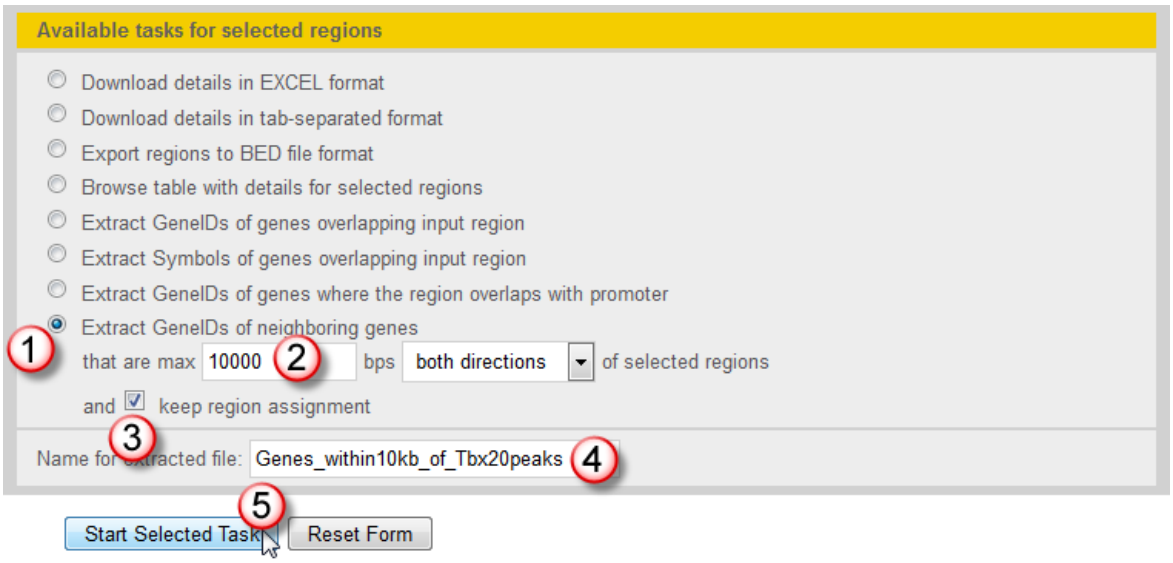
Select the option *Browse table with details…*, and start the task.



The output shows the neighboring gene loci for each region in both directions and on both strands, as well as overlaps with promoters, exons, and introns, and the number of V$BRAC binding site matches in each peak.



Next, please go back to the overview page, and select the option *Extract GeneIDs of neighboring genes*. For this example, set the maximum distance to 10,000 bp. To include the identifiers of the corresponding peaks, activate the *keep region assignment* option. Provide a file name, and save the file with the GeneIDs on your local computer.



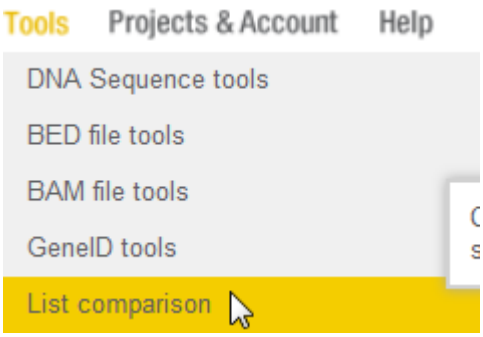
The file contains the GeneIDs and associated peak identifiers based on the peak IDs in the BED file.

11287	Region_1019					
11304	Region_571					
11426	Region_699					
11430	Region_1774	Region_1774				
11459	Region_1330					
11461	Region_884	Region_884	Region_885	Region_885	Region_886	Region_886
11464	Region_392	Region_393				
11465	Region_676					
11472	Region_1902					
11504	Region_2308					
11512	Region_2214	Region_2214				
11520	Region_659					
11539	Region_107					
11639	Region_674	Region_675				
11652	Region_1053					
11790	Region_54	Region_54				
11804	Region_1379					
11811	Region_2405					
11818	Region_1765					
11829	Region_2473					

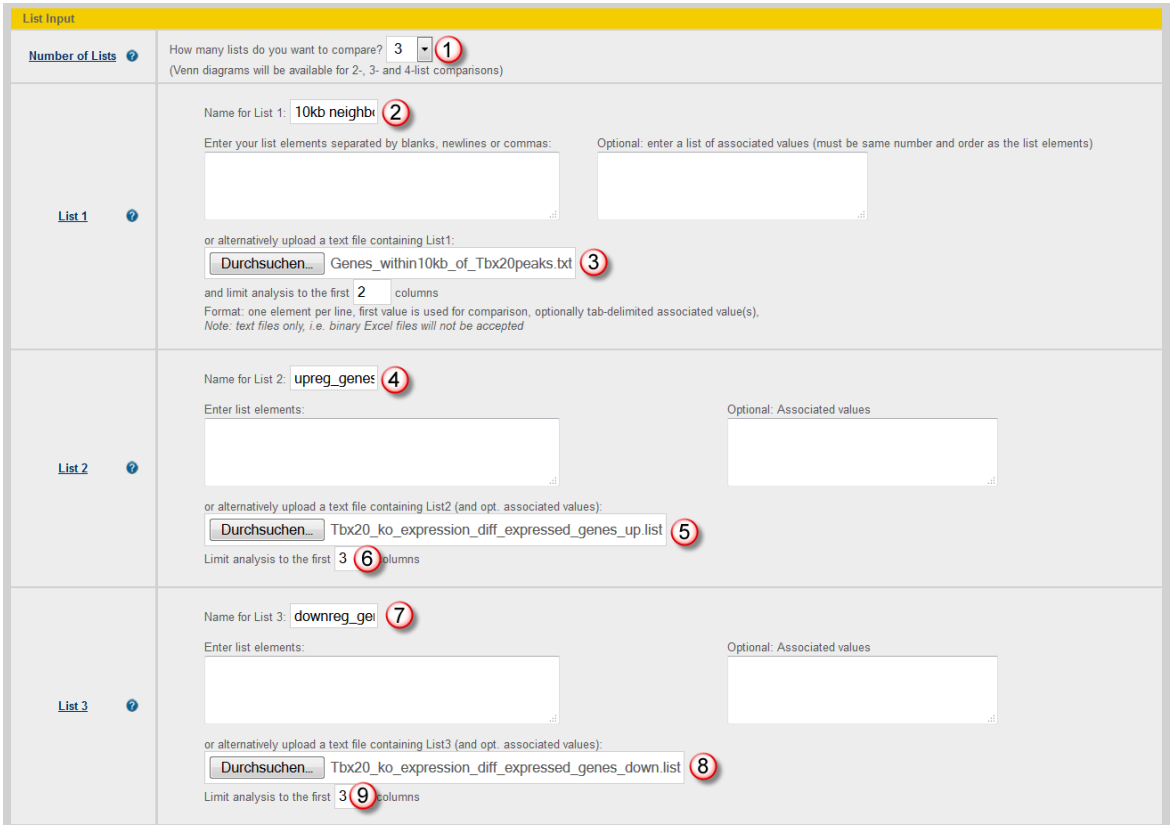
Comparison of Tbx20-neighboring genes with regulated genes

As we have expression data available, we can now compare the list of Tbx20-neighboring genes with the previously saved lists of up- and down-regulated genes in Tbx20 knock-out mouse hearts.

Start the *List comparison* tool from the *Tools* menu in the navigation bar.

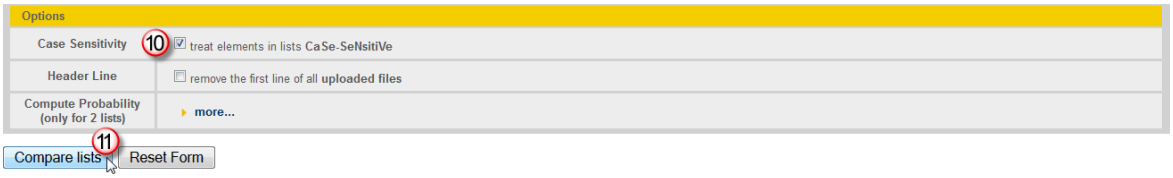


As you will compare three lists to one another, namely the list of Tbx20 neighboring genes the list of up-regulated genes, and the list of down-regulated genes from the expression analysis, set the number of lists accordingly to 3 (marked with 1 in the screenshot below). Provide a name for each list (2,4,7), and upload the corresponding files from your computer (3,5,8).

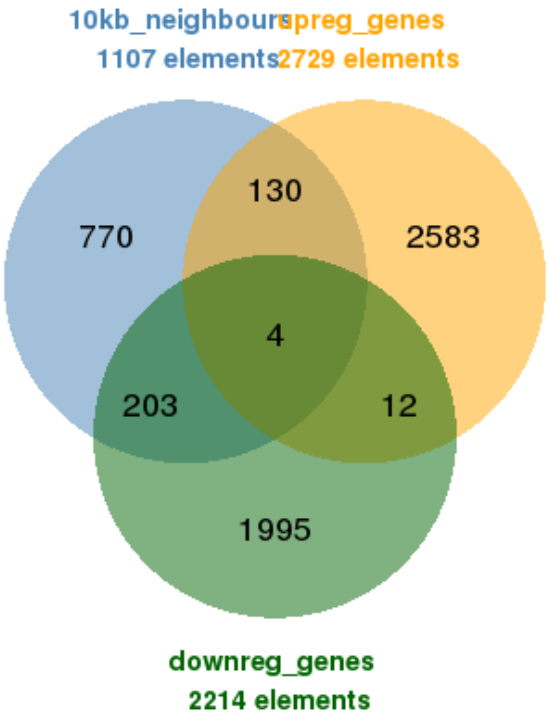


The list comparison tool allows to include associated values in the output. For uploaded tab-separated text files, you can select how many columns should be evaluated for each file. the default is 2, i.e. the identifier column plus one column with associated values. Set this value to 3 for the files with up- and down-regulated genes (6,9). This will included fold change values and gene names in addition to the gene IDs.

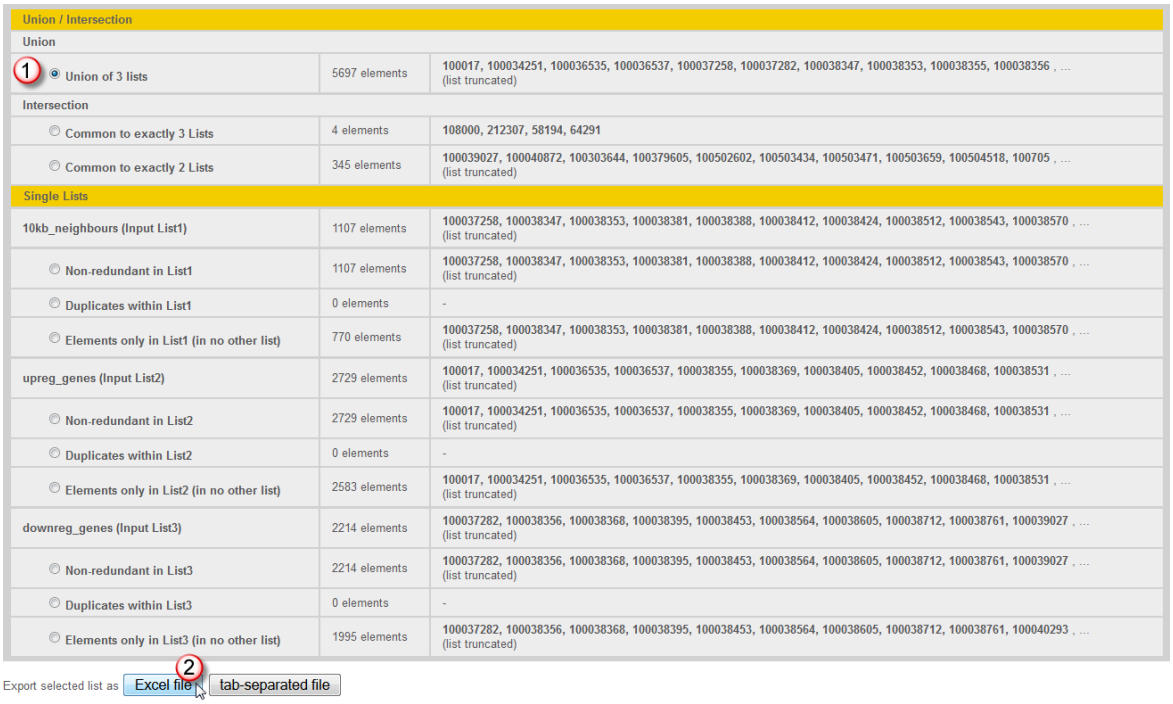
To keep the case as it is in the uploaded files, activate the *Case Sensitivity* option (otherwise lower case will be converted to upper case in the output). This also makes the ID comparison case-sensitive. The start the comparsion.



In the result, you'll find a Venn diagram with the overlap numbers. Of the 1107 neighboring genes, 134 are also found in the up-regulated list, and 207 in the down-regulated list.



To see the complete comparison, export the union of all lists to Excel.



Union / Intersection		
Union		
(1) Union of 3 lists	5697 elements	100017, 100034251, 100036535, 100036537, 100037258, 100037282, 100038347, 100038353, 100038355, 100038356 , ... (list truncated)
Intersection		
Common to exactly 3 Lists	4 elements	108000, 212307, 58194, 64291
Common to exactly 2 Lists	345 elements	100039027, 100040872, 100303644, 100379605, 100502602, 100503434, 100503471, 100503659, 100504518, 100705 , ... (list truncated)
Single Lists		
10kb_neighbours (Input List1)	1107 elements	100037258, 100038347, 100038353, 100038381, 100038388, 100038412, 100038424, 100038512, 100038543, 100038570 , ... (list truncated)
Non-redundant in List1	1107 elements	100037258, 100038347, 100038353, 100038381, 100038388, 100038412, 100038424, 100038512, 100038543, 100038570 , ... (list truncated)
Duplicates within List1	0 elements	-
Elements only in List1 (in no other list)	770 elements	100037258, 100038347, 100038353, 100038381, 100038388, 100038412, 100038424, 100038512, 100038543, 100038570 , ... (list truncated)
upreg_genes (Input List2)	2729 elements	100017, 100034251, 100036535, 100036537, 100038355, 100038369, 100038405, 100038452, 100038468, 100038531 , ... (list truncated)
Non-redundant in List2	2729 elements	100017, 100034251, 100036535, 100036537, 100038355, 100038369, 100038405, 100038452, 100038468, 100038531 , ... (list truncated)
Duplicates within List2	0 elements	-
Elements only in List2 (in no other list)	2583 elements	100017, 100034251, 100036535, 100036537, 100038355, 100038369, 100038405, 100038452, 100038468, 100038531 , ... (list truncated)
downreg_genes (Input List3)	2214 elements	100037282, 100038356, 100038368, 100038395, 100038453, 100038564, 100038605, 100038712, 100038761, 100039027 , ... (list truncated)
Non-redundant in List3	2214 elements	100037282, 100038356, 100038368, 100038395, 100038453, 100038564, 100038605, 100038712, 100038761, 100039027 , ... (list truncated)
Duplicates within List3	0 elements	-
Elements only in List3 (in no other list)	1995 elements	100037282, 100038356, 100038368, 100038395, 100038453, 100038564, 100038605, 100038712, 100038761, 100040293 , ... (list truncated)

Export selected list as (2) Excel file | tab-separated file

Genes that were present in each input list have an associated value (Region ID for neighboring genes; log fold change and gene symbol for regulated genes); the others get only a dash in the value columns.

Element	value(s) from 10kb neighbours	value(s) from upreg_genes		value(s) from downreg_genes	
100017	-	1.307	Ldlrap1	-	-
100034251	-	1.782	Wfdc17	-	-
100036535	-	2.338	Gm9913	-	-
100036537	-	2.263	Gm11149	-	-
100037258	Region_2099	-	-	-	-
100037282	-	-	-	-1.645	Rsph3b
100038347	Region_1097	-	-	-	-
100038353	Region_2543	-	-	-	-
100038355	-	2.291	Runx2os1	-	-
100038356	-	-	-	-1.313	Gm15612
100038368	-	-	-	-1.474	Gm10609
100038369	-	1.653	F630201L12Rik	-	-
100038381	Region_1416	-	-	-	-
100038388	Region_1318	-	-	-	-
100038395	-	-	-	-1.49	1700061E17Rik
100038405	-	3.872	Gm10827	-	-
100038412	Region_1195	-	-	-	-
100038424	Region_606	-	-	-	-
100038452	-	2.469	Gm13372	-	-
100038453	-	-	-	-2.452	Gm12522
100038468	-	2.654	Gm10684	-	-
100038512	Region_635	-	-	-	-
100038531	-	1.503	D030062O11Rik	-	-
100038543	Region_2235	-	-	-	-
100038548	-	2.36	Gm10521	-	-
100038564	-	-	-	-1.81	Gm10524
100038570	Region_1775	-	-	-	-
100038605	-	-	-	-1.512	E030047D23Rik
100038610	Region_1606	-	-	-	-

Thus you can use Excel functionality to filter e.g. for up-regulated Tbx20-neighboring genes (which would correspond to genes whose expression is probably directly repressed by Tbx20).

Element	value(s) from 10kb neighbours	value(s) from upreg_genes		value(s) from downreg_genes	
100379605	Region_1501	2.152	Gm15270	-	-
100503434	Region_2453	2.537	Gm19689	-	-
100503471	Region_176	1.943	Gm15867	-	-
100503659	Region_1562	1.453	Dos	-	-
102595	Region_1413	1.302	Plekho2	-	-
102631551	Region_2547	1.669	LOC102631551	-	-
105245	Region_1933	1.336	Txndc5	-	-
105988	Region_2224	2.486	Espl1	-	-
106205	Region_2236	1.071	Zc3h7a	-	-
107702	Region_1192	1.722	Rnh1	-	-
107765	Region_2582	2.879	Ankrd1	-	-
108000	Region_161	2.281	Cenpf	-1.399	Cenpf
108099	Region_773	1.074	Prkag2	-	-
108903	Region_1787	1.214	Tbcd	-	-
108912	Region_2059	2.517	Cdca2	-	-
11304	Region_571	1.87	Abca4	-	-
11459	Region_1330	1.127	Acta1	-	-
11461	Region_884	1.774	Actb	-	-
11465	Region_676	1.771	Actg1	-	-
11504	Region_2308	2.044	Adamts1	-	-
11520	Region_659	1.241	Plin2	-	-
12181	Region_2177	1.276	Bop1	-	-
12523	Region_147	3.191	Cd84	-	-
12606	Region_1057	1.915	Cebpa	-	-
12982	Region_2617	1.834	Csf2ra	-	-
14087	Region_1329	2.129	Fanca	-	-

The identifiers can then, for example, be uploaded to the Genomatix Pathway System for further analysis.

Literature

Audic S, Claverie JM. The significance of digital gene expression profiles. Genome Res 10, 986-995 (1997).

Sakabe NJ, Aneas I, Shen T, Shokri L et al. Dual transcriptional activator and repressor roles of TBX20 regulate adult cardiac structure and function. Hum Mol Genet 21(10), 2194-2204 (2012).

Stacklies W, Redestig H, Scholz M, Walther D, Selbig J: pcaMethods - a Bioconductor package providing PCA methods for incomplete data. Bioinformatics 23, 1164-1167 (2007).

Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W: A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics 25, 1952-1958 (2009)

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, Liu XS: Model-based analysis of ChIP-Seq (MACS). Genome Biol 9(9), R137 (2008).

List of resources available on the web:

Gene Expression Omnibus:
http://www.ncbi.nlm.nih.gov/geo/

Further reading:
http://www.genomatix.de/expertise/publications.html

This tutorial was compiled for Genomatix Genome Analyzer v3.51106.

Please note that depending on the program versions and database releases used slight variations in results (e.g. gene numbers) may occur.